Clinical Research Informatics: Contributions from 2016

C. Daniel1,2, R. Choquet2, Section Editors for the IMIA Yearbook Section on Clinical Research Informatics
1 AP-HP Direction of Information Systems, Paris, France
2 INSERM UMRS 1142, Paris, France

Summary

Objectives: To summarize key contributions to current research in the field of Clinical Research Informatics (CRI) and to select the best papers published in 2016.

Methods: A bibliographic search using a combination of MeSH and free terms on CRI was performed using PubMed, followed by a double-blind review in order to select a list of candidate best papers to be then peer-reviewed by external reviewers. A consensus meeting between the two section editors and the editorial team was organized to finally conclude on the selection of best papers.

Results: Among the 452 papers published in 2016 in the various areas of CRI and returned by the query, the full review process selected four best papers. The authors of the first paper utilized a comprehensive representation of the patient medical record and semi-automatically labeled training sets to create phenotype models via a machine learning process. The second selected paper describes an open source tool chain securely connecting ResearchKit compatible applications (Apps) to the widely-used models via a machine learning process. The third selected paper describes the FAIR Guiding Principles for scientific data management and stewardship. The fourth selected paper focuses on the evaluation of healthcare-generated data, person health data, and patient-reported outcomes.

Keywords
Clinical Research Informatics; biomedical research; real-world data; phenotyping; data integration

Introduction

The goal of this section is to provide an overview of research trends from 2016 publications that demonstrate excellent research about the multifaceted aspects of medical informatics supporting clinical trials and observational studies. Clinical Research Informatics (CRI) continues to be developed and CRI community has especially to address the important challenges related to “Learning from experience and secondary use of patient data” - this year’s special topic of the IMIA Yearbook. New methods and tool chains have been developed in order to collect, integrate, and mine “real-world data” – healthcare-generated data, person health data and patient-reported outcomes.

About the Paper Selection

A comprehensive review of articles published in 2016 addressing a wide range of issues for CRI was conducted. The selection was performed by querying MEDLINE via PubMed (from NCBI, National Center for Biotechnology Information) with a set of predefined MeSH descriptors: Biomedical Research, Clinical research, Medical research, Pharmacovigilance, Patient Selection, Phenotyping, Genotype-phenotype associations, Data Collection, Epidemiologic Research Design, Epidemiologic Study Characteristics as Topic, Epidemiologic Monitoring, Evaluation Studies as Topic, Clinical Trials as Topic, Feasibility Studies. References addressing topics of other sections of the Yearbook, such as Translational Bioinformatics, were excluded based on predefined exclusion MeSH descriptors such as Genetic Research, Gene Ontology, Human Genome Project, Stem Cell Research, or Molecular Epidemiology.

Bibliographic databases were searched on January 27, 2017 for papers published in 2016, considering the electronic publication date. From an original set of 906 references, a first subset of 452 references was considered according to its relevancy to the CRI field and blindly reviewed by the two section editors based on papers’ title and abstract. The articles were classified into several CRI categories: i) CRI for clinical trials, observational studies, and real-world data; ii) data management (data collection and integration, data quality, open data); iii) data mining and machine learning techniques; iv) data privacy, security and regulatory issues, and v) policy and patient perspectives. Their contribution to CRI was rated as low, medium or high. Then, the two lists of references were merged, yielding 170 references classified as “high contribution” to CRI by at least one reviewer or as “medium contribution” by both reviewers. The 170 references were reviewed jointly by the two section editors to select a consensual list of 16 candidate best papers representative of all CRI categories. Following the IMIA Yearbook process, these candidate best papers were peer-reviewed by editors and external reviewers (at least four reviewers per paper). Four papers were finally selected as best papers (Table 1). A content summary of these selected papers can be found in the appendix of this synopsis.
Conclusions and Outlook

CRI for Observational Studies and Real World Data

Healthcare-generated data has become an important resource for clinical and genomic research. Often, investigators create and iteratively refine phenotype algorithms to achieve high positive predictive values or sensitivity, thereby identifying valid cases and controls. Kirby et al. [1] reported the current status and impact of the Phenotype Knowledge Base (PheKB, http://phekb.org), an online environment supporting the workflow of building, sharing, and validating electronic phenotype algorithms, and they demonstrated that a broad range of algorithms used to mine electronic health record data from different health systems, and generally transportable across the sites, have significantly high performance.

Machine learning approaches running on real-world data are limited by the paucity of labeled training datasets. Traditionally, patient groups with a given phenotype are selected through rule-based definitions (see PheKB initiative) whose creation and validation are time-consuming. The first selected paper by Agarwal et al. addresses the limitation of the generation of clinical phenotype descriptions. Using the Halpern et al. method based on “anchor” terms [2], the authors demonstrated the feasibility of utilizing semi-automatically labeled training sets to create phenotype models via machine learning, using a comprehensive representation of the patient medical record [3]. They validated the phenotype models in the context of Type 2 diabetes mellitus (T2DM) and Myocardial Infarcts (MI) using respectively the phenotype definitions of the eMERGE [1] and OMOP [4] initiatives. Similarly, by combining de-noising auto-encoders with random forests, Beaulieu et al. [5] found classification improvements across multiple simulation models and improved survival prediction in amyotrophic lateral sclerosis (ALS) clinical trial data. Such approaches can accelerate research with large observational healthcare datasets.

Personal health data and Patient Reported Outcomes (PROs) are also “real-world data” and have the most value when presented in context along with health system data. With the transformation of smartphones into personal health data storage devices, there is a need to provide data transmission facilities and to connect research Apps to the health system. The second selected paper from Pfiffner et al. describes C3-PRO (Consent, Contact, and Community framework for Patient Reported Outcomes), an open source tool chain securely connecting, in a standards-compliant fashion, ResearchKit compatible Apps to the clinical research infrastructure Informatics for Integrating Biology and the Bedside (i2b2), widely adopted by 140 academic medical centers [6]. The case study from Harle et al. [7] describes a novel information system for electronic collection of Patient-Reported Outcomes (PRO) and the lessons learned in implementing that system to support research in an academic health center [5].

Data Collection and Integration

Luo et al. proposed an hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system [8]. Common data models (CDMs) need to be built for sharing data from large, longitudinal, Electronic Health Record (EHR)-based community registries. However, each new data research network that wishes to support its own analytics tends to develop its own data model. Gaya et al. evaluated four CDMs in use for clinical research data: Sentinel v5.0 (referred to as the Mini-Sentinel CDM in previous versions), PCORnet v3.0 (an extension of the Mini-Sentinel CDM), OMOP v5.0, and CDISC SDTM v1.4 [9]. Klann et al. proposed an approach using i2b2 as a hub, to rapidly reconfigure data to meet new analytical requirements without new Extracting Transforming and Loading (ETL) programming and evaluated this approach to generate a PCORnet Common Data Model physical database from existing i2b2 systems [10]. There are limited toolboxes enabling the creation of reusable and machine-executable phenotype algorithms, which has hampered effective cross-institutional research collaborations. Jiang et al. developed and evaluated a data element repository (DER) for providing machine-readable data element service Application Programming Interfaces (APIs) to support phenotype algorithm authoring and execution [11]. Anguita et al. proposed a method and software framework for enriching private biomedical sources with data from public online repositories [12].

Data Quality

Johnson et al. [13] applied an ontology-based assessment process to EHR...
data and determined its usefulness in characterizing data quality for calculating an example eMeasure [11]. Bruland et al. evaluated the completeness of EHR data for secondary uses of routinely collected patient data [14].

Open Data

Current digital ecosystem surrounding scholarly data publication still prevents us from extracting maximum benefit from research investments. The third selected paper from Wilkinson et al. describes the FAIR Guiding Principles for scientific data management and stewardship [15]. This concise and measurable set of principles may act as guidelines for those wishing to enhance the reusability of their data holdings. The FAIR Guiding Principles put a specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.

Data Privacy

Given the growing list of quasi identifiers in molecular phenotype datasets and potentially linkable datasets, the risk of different types of privacy breaches must be considered. The fourth selected paper from Harmanci et al. focuses on the evaluation of the risk of privacy breaches in releasing genomics datasets [16]. The authors investigated how far molecular phenotype data (such as gene expression level) can be - in contrast to DNA variants - considered as free of identifying information as it is generally assumed. They proposed a framework for practical instantiation of linking attacks using a genotyping dataset and publicly available anonymized phenotype datasets and genotype-phenotype correlations. The authors proposed statistical quantification methods to objectively quantify the risk of linking attacks before releasing a genotyping dataset. The methods proposed by the authors can be integrated into the existing risk assessment and management strategies.

Policy and Patient Perspective

More generally speaking, in the wake of public and policy concerns about security and inappropriate use of data, conventional approaches toward data governance may no longer be sufficient to respect and protect individual privacy. One proposed solution to improve transparency and public trust is known as the Dynamic Consent, which uses information technology to facilitate a more explicit and accessible opportunity to opt out. Spencer et al. evaluated the patient perceptions of a dynamic consent model and electronic system to enable and implement ongoing communication and collaboration between patients and researchers [17]. Patients from a range of socioeconomic backgrounds viewed a digital system for dynamic consent positively, in particular, feedback about data recipients and research results.

In conclusion, a major trend in the 2016 publications concerns the variety of research on “real-world data” - healthcare-generated data, person health data, and patient-reported outcomes - highlighting opportunities provided by new machine learning techniques as well as new potential risks of privacy breaches.

Acknowledgement

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

References

4. OMOP. (2016) OMOP Health Outcomes of Interest: Observational Medical Outcomes Partnership [Internet]. [cité 30 mai 2017]. Available at: http://omop.org/HOI