Clinical Research Informatics: Contributions from 2016

C. Daniel[1,2], R. Choquet[2], 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 clinical research infrastructure Informatics for Integrating Biology and the Bedside (i2b2). The third selected paper describes the FAIR Guiding Principles for scientific data management and stewardship. The fourth selected paper focuses on the evaluation of the risk of privacy breaches in releasing genomics datasets.

Conclusions: A major trend in the 2016 publications is the variety of research on "real-world data" - healthcare-generated data, person health data, and patient-reported outcomes – highlighting the opportunities provided by new machine learning techniques as well as new potential risks of privacy breaches.

Keywords

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

Yearb Med Inform 2017:209-13
http://dx.doi.org/10.15265/IY-2017-024
Published online August 18, 2017

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, Epidemiological 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].

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2017 in the section 'Clinical Research Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Clinical Research Informatics
▪ Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, Sweeney TE, Gyang E, Shah NH. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc 2016;23(6):1166-73.
▪ Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat Methods 2016;13(3):251-6.
▪ Pfiffner PB, Pinyol I, Natter MD, Mandl KD. C3-PRO: Connecting ResearchKit to the Health System Using i2b2 and FHIR. PloS One 2016;11(3):e0152722.
▪ Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.

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. Gaza *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 measureable 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 genotype 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 genotype 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

1. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc 2016;23(6):1046-52.
2. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc 2016;23(4):731-40.
3. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc 2016;23(6):1166-73.
4. OMOP. Health Outcomes of Interest | Observational Medical Outcomes Partnership [Internet]. [cité 30 mai 2017]. Available at: http://omop.org/HOI
5. Beaulieu-Jones BK, Greene CS, Pooled Resource Open-Access ALS Clinical Trials Consortium. Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform 2016;64:168-78.
6. Pfiffner PB, Pinyol I, Natter MD, Mandl KD. C3-PRO: Connecting ResearchKit to the Health System Using i2b2 and FHIR. PloS One 2016;11(3):e0152722.
7. Harle CA, Lipori G, Hurley RW. Collecting, Integrating, and Disseminating Patient-Reported Outcomes for Research in a Learning Healthcare System. EGEMS (Wash DC) 2016;4(1):1240.
8. Luo L, Li L, Hu J, Wang X, Hou B, Zhang T, et al. A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. BMC Med Inform Decis Mak 2016;16:114.
9. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016;64:333-41.
10. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc 2016;23(5):909-15.
11. Jiang G, Kiefer RC, Rasmussen LV, Solbrig HR, Mo H, Pacheco JA, et al. Developing a data element repository to support EHR-driven phenotype algorithm authoring and execution. J Biomed Inform 2016;62:232-42.
12. Anguita A, García-Remesal M, Graf N, Maojo V. A method and software framework for enriching private biomedical sources with data from public online repositories. J Biomed Inform 2016;60:177-86.
13. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Application of An Ontology for Characterizing Data Quality For a Secondary Use of EHR Data. Appl Clin Inform 2016;7(1):69-88.
14. Bruland P, McGilchrist M, Zapletal E, Acosta D, Proeve J, Askin S, et al. Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. BMC Med Res Methodol 2016;16(1):159.
15. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
16. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat Methods 2016;13(3):251-6.
17. Spencer K, Sanders C, Whitley EA, Lund D, Kaye J, Dixon WG. Patient Perspectives on Sharing Anonymized Personal Health Data Using a Digital System for Dynamic Consent and Research Feedback: A Qualitative Study. J Med Internet Res 2016;18(4):e66.

Correspondence to:
Christel Daniel, MD, PhD
WIND DSI – Assistance Publique – Hôpitaux de Paris
5 rue Santerre
75 012 Paris, France
Tel: +33 1 48 04 20 29
E-mail: christel.daniel@aphp.fr

Summary of Best Papers Selected for the IMIA Yearbook 2016, Section Clinical Research Informatics

Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, Sweeney TE, Gyang E, Shah NH

Learning statistical models of phenotypes using noisy labeled training data

J Am Med Inform Assoc 2016;23(6):1166-73

Machine learning approaches running on real-world data are limited by the paucity of labeled training datasets. Traditionally, patient groups sharing a given phenotype are selected through rule-based definitions which creation and validation are time-consuming. This paper addresses the limitation of the generation of clinical phenotype descriptions and demonstrates 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. The authors provide an extended background about i) manual rule-based definition of phenotypes for research purposes; ii) learning techniques (based on Natural Language Processing and/or other techniques) using manually created training sets (labeled cases and controls built from chart review), and iii) noise tolerant learning techniques.

Based on the phenotype definitions provided by the eMERGE [1] and OMOP [4] initiatives, the authors automatically identified within the Stanford Clinical Data Repository 32,581 possible cases for T2DM and 36,858 possible cases for MI. Using the Halpern *et al.* method based on "anchor" terms [2], they defined a list of keywords specific to the phenotypes of interest to semi-automatically generate noisy labeled training data. Then, a sample of 1,500 patient records - 750 patient records for each phenotype having a "noisy" label for the phenotype and 750 controls taken in the extract disjoint with possible cases (silver standard) - was used to train the XPRESS (eXtraction of Phenotypes from Records using Silver Standards) model. The building of XPRESS models consisted of feature engineering from structured and unstructured data and learning statistical models from the noisy labeled data. The performance of the models was evaluated against a gold standard consisting of a clinician-reviewed evaluation set (gold standard: cases and controls created by five clinicians, disjoint from the records used for the training). The models for T2DM and MI achieved a precision and accuracy of 0.90, 0.89, and 0.86, 0.89, respectively. Local implementations of the previously validated rule-based definitions for T2DM and MI achieved precision and accuracy of 0.96, 0.92 and 0.84, 0.87, respectively. The authors demonstrated that they can learn phenotype models of chronic and acute phenotypes from 4,135 noisy labeled training samples (XPRESS models) acquired at a negligible cost with the same performance as from 2,026 manually labeled, zero-error training samples. Using imperfectly labeled data, the method provides an alternative to manual labeling for creating training sets. Such an approach may be used to create local phenotype models and can accelerate research with large observational healthcare datasets. Further research in feature engineering and in the specification of the keyword list can improve the performance of the models and the scalability of the approach.

Pfiffner PB, Pinyol I, Natter MD, Mandl KD

C3-PRO: Connecting ResearchKit to the Health System Using i2b2 and FHIR

PloS One 2016;11(3):e0152722

As new mobile technologies are more widely adopted, their use for care and research is being more and more efficient. One of the actual challenges is to connect research Apps to the healthcare system and use real life patient-generated data in order to improve pharmacovigilance and to obtain medication observance data for post market studies or other usages. In March 2015, Apple Inc. deployed a new open source framework to help research promoters to build easy smartphones Apps for clinical studies. To complete the system, the authors extended the Apple ResearchKit with a Consent, Contact, and Community framework for Patient Reported Outcomes (namely C3-PRO). The aim of this extension is to connect the ResearchKit App to the widely used clinical research IT infrastructure i2b2. C3-PRO enables a method to create eligibility criteria question, informed consent, and participant surveys using FHIR data formats. Data is encrypted prior to be sent over the Internet. It is then pushed into an i2b2 FHIR compatible cell. The paper describes the complete system and the data flows including security measures both in terms of data processing and at the App level. The system can collect data anonymously, using the UUID (Universally Unique Identifier) of the device as identifier. The system can also capture sensor-based data. Using the system, recruitment for studies can be done more widely and faster. The resulting data processing is then taking advantage of i2b2 generic architecture to process classic statistics and produce first reports. Besides, authors are working on mechanisms for data-linkage with existing cohorts as well as a cross platform version or their kit (Android/Iphone). By leveraging the FHIR formats, C3-PRO enables survey question and consent libraries to become standardized and used across studies.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B

The FAIR Guiding Principles for scientific data management and stewardship

Sci Data 2016 Mar 15;3:160018

The current digital ecosystem of scholarly data publication still prevents us from extracting the maximum benefit from research

investments. Science funders, publishers, and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Main barriers to data reusability are not technical. An appropriate set of basic principles to data stewardship to be followed by database owners, data managers, or data scientists is proposed in this paper to integrate and propagate digital object "best design" rules. The authors present four foundational principles, the FAIR guiding principles, that are related but independent and separable. The FAIR guiding principles are setting basic rules so that data should be: Findable, Accessible, Interoperable and Reusable. Operational rules are defined for each principle. For instance, to be *Findable*, a unique and persistent identifier should be assigned to data and metadata, rich metadata should describe data, and metadata and data should be registered and indexed in a searchable source. To be *Accessible*, (meta) data are retrievable by their identifier using a standard, open, and free protocol allowing authentication and authorization procedures, and allowing metadata to be accessible even when data is not. To be *Interoperable*, (meta)data should use a formal, accessible, and shared set of broadly applicable languages and vocabularies for knowledge representation. And finally, to be *Reusable* (meta) data should be richly described with a plurality of attributes and be released with provenance and clear licensing information. These principles do not suggest for any specific technology, nor standard or implementation-solution. Many scientific datasets or projects, such as Dataverse, FAIRDOM, ISA, Open PHACTS, or UniProt, are already implementing some of these principles. Although FAIR principles are not a technical standard, they put a specific emphasis on enhancing the ability of machines to automatically find and use data, in addition to supporting its reuse by individuals.

Harmanci A, Gerstein M

Quantification of private information leakage from phenotype-genotype data: linking attacks

Nat Methods 2016 Mar;13(3):251-6

As the number and size of phenotype and genotype datasets increase, the privacy protection of individuals is emerging as an important issue. This paper focuses on the evaluation of the risk of privacy breaches in releasing genomics datasets. Harmanci *et al.* 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.

The authors provide a background about the growing list of quasi identifiers in molecular phenotype datasets and about two different types of privacy breaches. These privacy breaches result from either detecting whether an individual with known genome has participated to a study or cross-referencing of multiple seemingly independent genotype and phenotype datasets (knowing that the number of potentially linkable datasets will increase). They propose a framework for practical instantiation of linking attacks using a genotype dataset and publicly available anonymized phenotype datasets and genotype-phenotype correlations. The authors emphasize the need of statistical quantification methods to objectively quantify the risk of linking attacks before releasing a genotype dataset. They propose two measures: the cumulative individual characterization information (ICI) and the genotype predictability. ICI is described as the total amount of information in a set of variant genotypes that can be used in a linking attack. For a set of variants, genotype predictability measures how predictable genotypes are, given the gene expression levels. A three-step framework for instantiating linking attacks is presented. Based on the framework implementation on a test set, authors demonstrated that more than 95% of individuals are vulnerable and they observed that the extremity attacks (extreme of the gene expression levels observed with extremes of the phenotypes) can link family members within the dataset. Once the risk assessment is performed, several strategies can be set to minimize risks. For example k-anonymization proposes to censor entries or adding noise into the dataset on specific data points that have been characterized as possible leaks (ICI). Finally, inclusion of biological utility measures should be done along with the risk assessment. The methods proposed by the authors can be integrated directly into the existing risk assessment and management strategies.