Published online: 2020-08-21

Clinical Research Informatics

Christel Daniel[1,2], Dipak Kalra[3], Section Editors for the IMIA Yearbook Section on Clinical Research Informatics

[1] Information Technology Department, AP-HP, Paris, France
[2] Sorbonne University, University Paris 13, Sorbonne Paris Cité, INSERM UMR_S 1142, LIMICS, Paris, France
[3] The University of Gent, Gent, Belgium

Summary

Objectives: To summarize key contributions to current research in the field of Clinical Research Informatics (CRI) and to select best papers published in 2019.

Method: A bibliographic search using a combination of MeSH descriptors and free-text 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. After peer-review ranking, a consensus meeting between the two section editors and the editorial team was organized to finally conclude on the selected three best papers.

Results: Among the 517 papers, published in 2019, returned by the search, that were in the scope of the various areas of CRI, the full review process selected three best papers. The first best paper describes the use of a homomorphic encryption technique to enable federated analysis of real-world data while complying more easily with data protection requirements. The authors of the second best paper demonstrate the evidence value of federated data networks reporting a large real world data study related to the first line treatment for hypertension. The third best paper reports the migration of the US Food and Drug Administration (FDA) adverse event reporting system database to the OMOP common data model. This work opens the combined analysis of both spontaneous reporting system and electronic health record (EHR) data for pharmacovigilance.

Conclusions: The most significant research efforts in the CRI field are currently focusing on real world evidence generation and especially the reuse of EHR data. With the progress achieved this year in the areas of phenotyping, data integration, semantic interoperability, and data quality assessment, real world data is becoming more accessible and reusable. High quality data sets are key assets not only for large scale observational studies or for changing the way clinical trials are conducted but also for developing or evaluating artificial intelligence algorithms guiding clinical decision for more personalized care. And lastly, security and confidentiality, ethical and regulatory issues, and more generally speaking data governance are still active research areas this year.

Keywords

Clinical research informatics; biomedical research, clinical trials as topic; observational studies as topic; real-world data; real-world evidence generation; phenotyping

Yearb Med Inform 2020:203-7
http://dx.doi.org/10.1055/s-0040-1702007

Introduction

For the 2020 International Medical Informatics Association (IMIA) Yearbook, the goal of the Clinical Research Informatics (CRI) section is to provide an overview of research trends from 2019 publications that demonstrate the progress in multifaceted aspects of medical informatics supporting research and innovation in the healthcare domain. New methods, tools, and CRI systems have been developed in order to enable real-world evidence generation and optimize the lifecycle of clinical trials. The CRI community has also addressed the important challenges of public trust in research in the era of Big Data and Artificial Intelligence contributing with studies related to "Ethics in Healthcare" - this year's special theme for the IMIA Yearbook.

Paper Selection Method

A comprehensive review of articles published in 2019 and 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 and free terms: *Clinical research informatics, Biomedical research, Nursing research, Clinical research, Medical research, Pharmacovigilance, Patient selection, Phenotyping, Genotype-phenotype associations, Feasibility studies, Eligibility criteria, Feasibility criteria, Cohort selection, Patient recruitment, Clinical trial eligibility screening, Eligibility determination, Patient-trial matching, Protocol feasibility, Real world evidence, Data Collection, Epidemiologic research design, Clinical studies as Topic, Multicenter studies as Topic, and Evaluation studies as Topic.* Papers addressing topics of other sections of the Yearbook, such as Translational Bioinformatics, were excluded based on the predefined exclusion of MeSH descriptors such as *Genetic research, Gene ontology, Human genome project, Stem cell research,* or *Molecular epidemiology*.

Bibliographic databases were searched on January 30, 2020 for papers published in 2019, considering the electronic publication date. Among an original set of 685 references, 517 papers were selected as being in the scope of CRI and their scientific quality was blindly rated as low, medium, or high by the two section editors based on papers' title and abstract. Eighty-three references classified as high quality contributions to the field by at least one of the section editors or as medium quality contributions by both of them were considered. These 83 papers were classified into the following eleven areas of the CRI domain in order of the number of matching papers (multiple classification choices were permitted): observational studies, reuse of electronic health record (EHR) data (n=48); data/text mining and algorithms (n=42); feasibility studies, patient recruitment, data management and CRI systems (n=18); ethical, legal, social, policy issues and solutions, stakeholder participation (n=15); data integration and semantic interop-

erability (n=12); data quality assessment or validation (n=9); security and confidentiality (n=4); and communicating study results (n=4). The 83 references were reviewed jointly by the two section editors to select a consensual list of 15 candidate best papers representative of all CRI categories. In conformance with the IMIA Yearbook process, these 15 papers were peer-reviewed by the IMIA Yearbook editors and external reviewers (at least four reviewers per paper). Three papers were finally selected as best papers (Table 1). A content summary of these best papers can be found in the appendix of this synopsis.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2020 in the section 'Clinical Research Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Clinical Research Informatics
▪ Paddock S, Abedtash H, Zummo J, Thomas S. Proof-of-concept study: Homomorphically encrypted data can support real-time learning in personalized cancer medicine. BMC Med Inform Decis Mak 2019 Dec 4;19(1):255.
▪ Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, Reich CG, Duke J, Madigan D, Hripcsak G, Ryan PB. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet 2019 Nov 16;394(10211):1816-26.
▪ Yu Y, Ruddy KJ, Hong N, Tsuji S, Wen A, Shah ND, Jiang G. ADEpedia-on-OHDSI: A next generation pharmacovigilance signal detection platform using the OHDSI common data model. J Biomed Inform 2019 Mar;91:103119.

Conclusions and Outlook

The 15 candidate best papers for 2020 illustrate recent efforts and trends in different CRI areas such as real-world evidence generation; data/text mining, Artificial Intelligence (AI) and Machine Learning (ML); feasibility studies, patient recruitment, data management and CRI systems; ethical, legal, social, policy issues and solutions, stakeholder participation; data integration, semantic interoperability and data quality assessment; and security and confidentiality.

Real-world Evidence Generation

If randomized clinical trials remain the reference methodology for biomedical research in terms of level of evidence, real-world data (RWD) is increasingly used to generate new knowledge. The first step of real-world evidence (RWE) generation consists of cohort building i.e., patient selection based on criteria that identify patients with a given condition or disease (phenotyping). CALIBER is a nationwide cardiovascular data initiative in the UK leveraging best practices from leading consortia (e.g., eMERGE, Million Veteran Program, BioVU) for developing, validating, and sharing reproducible phenotypes [1]. The **best paper** from Suchard *et al*., describes how the Observational Health Data Science and Informatics (OHDSI) distributed data network supported a large-scale observational study enabling effectiveness and safety evaluation of first-line drugs for hypertension [2]. The authors conclude the superiority of thiazide or thiazide-like diuretics to ACE inhibitors with reduced hospitalization for myocardial infraction, heart failure, and stroke and the inferiority of non-dihydropyridine calcium channel blockers. Wang *et al.,* provide a Baysian non-parametric causal inference model to address the challenge of synthesizing information from both clinical trials and registry studies for evaluating the causal effect of medical interventions and more generally healthcare decision making [3].

Data Integration, Semantic Interoperability and Data Quality Assessment

Data integration relies on the definition of common data model and vocabularies. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) has been adopted for standardizing the data from a variety of EHR data bases allowing systematic analysis of disparate observational databases within the OHDSI network. The **best paper** from Yu et al. reports the migration of the US Food and Drug Administration (FDA) adverse event reporting system database to the OMOP CDM and investigates the information loss of this transformation [4]. With the growing adoption of the OMOP CDM in hospitals, the resulting ADEpedia-on-OHDSI represents an integrative framework facilitating the combined analysis of both spontaneous reporting system and EHR data for pharmacovigilance. In order to facilitate the analysis of OMOP-formatted EHR data for R users, Glicksberg *et al.,* provide a package called ROMOP for data exploration, cohort building, and extraction of the data of the cohort patients [5].

Security and Confidentiality

RWD studies are hampered by data protection requirements. Data integration for modern data-driven research requires the development of privacy enhancing ETL (Extract, Transfer and Load) processes [6]. The best paper from Paddock et al., describes the use of a homomorphic encryption technique to enable the analysis of patient-level data whilst it remains in an encrypted state [7]. They demonstrated the feasibility of the approach in the context of drug repurposing in cancer using simulated patients and conventionally available computing facilities. This study offers a valuable method for limiting the risk of data re-identification in RWD studies, therefore complying more easily with data protection requirements. Synthetic clinical data is also a promising solution for exploring clinical data while protecting patient confidentiality. Chen et al., used clinical quality measures to investigate the capacity of the Synthea™ tool to generate realistic clinical data [8].

Data/Text Mining, Artificial Intelligence, and Machine Learning

In the era of big data and new technologies (AI/ML), RWD is increasingly used in order to develop innovative applications and new services supporting disease diagnosis, out-

come prediction, or therapeutic decision. Increasingly varied datasets are used to assess disease risk at an individual level, detect preclinical conditions, and initiate preventive strategy. Using omics and wearable monitoring data, Shussler *et al.,* developed molecular and physiological profiling of patient at risk of type 2 diabetes mellitus and developed prediction model for insulin resistance [9]. One of the challenges in the design and development of AI systems in conjunction with EHRs is to identity the possible biases and to remediate them. Fang *et al.,* investigated bias in applying ML to predict individual treatment effects [10].

Recent methods for patient profiling combine more and more frequently rule-based approaches and AI/ML-based models making use of clinical text, which remains one of the most important sources of phenotype. Zhang *et al.,* provide a detailed description of the PheCAP protocol, a high-throughput semi-supervised pipeline for phenotype generation using structured data and information extracted from the narrative notes [11]. Clinical trial recruitment is often the driving requirement for phenotyping. Meystre *et al*., evaluated different methods – pattern matching (regular expressions), ML-based Natural language processing (NLP) – to develop an automatic trial eligibility surveillance based on unstructured clinical data in breast cancer domain [12].

Feasibility Studies, Patient Recruitment, Data Management, and CRI Systems

Another type of innovation in clinical trial recruitment is to combine an EHR-based recruitment system with secure messaging used to contact and recruit patent directly [13]. Upstream, at the stage of protocol optimization, Claerhout *et al.,* demonstrated the value of distributed research networks for estimating at large scale the number of patients matching eligibility criteria [14]. They also stressed out the need to increase both phenotyping capabilities at hospital side and eligibility criteria editing at clinical research side. At the step of protocol execution, Carrigan *et al.,* investigated the use of EHRs to derive control arms for early phase single arm cancer trials [15].

Ethical, Legal, Social, Policy Issues and Solutions, Stakeholder Participation

Ethical standards and public trust in clinical research are major issues as illustrated by the choice of "Ethics in Healthcare" as the special topic of the 2020 edition of the Yearbook and recent studies are focusing on this topic. Beier *et al.,* investigated the concept of patient participation in data-driven research involving the linkage of massive heterogeneous data *e.g*., demographic, clinical routine, research and patient-reported data and also non-medical social or environmental data [16]. They observed that an inflationary use of participatory rhetoric can undermine public trust in data-driven research initiatives and concluded that education programs and communicative skills for deliberation should be considered within research plans and budgets.

Acknowledgement

We would like to acknowledge the support of Adrien Ugon, Martina Hutter, Kate Fultz Hollis, Lina Soualmia, Brigitte Séroussi, and the whole Yearbook editorial team as well as the reviewers for their contribution to the selection process of the Clinical Research Informatics section of the IMIA Yearbook 2020.

References

1. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc 2019;26(12):1545–59.
2. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet 2019;394(10211):1816–26.
3. Wang C, Rosner GL. A Bayesian nonparametric causal inference model for synthesizing randomized clinical trial and real-world evidence. Stat Med 2019;38(14):2573–88.
4. Yu Y, Ruddy KJ, Hong N, Tsuji S, Wen A, Shah ND, Jiang G. ADEpedia-on-OHDSI: A next generation pharmacovigilance signal detection platform using the OHDSI common data model. J Biomed Inform 2019;91:103119.
5. Glicksberg BS, Oskotsky B, Giangreco N, Thangaraj PM, Rudrapatna V, Datta D, et al. ROMOP: a light-weight R package for interfacing with OMOP-formatted electronic health record data. JAMIA Open 2019;2(1):10–4.
6. Prasser F, Kohlbacher O, Mansmann U, Bauer B, Kuhn KA. Data Integration for Future Medicine (DIFUTURE). Methods Inf Med 2018;57(S01):e57–e65.
7. Paddock S, Abedtash H, Zummo J, Thomas S. Proof-of-concept study: Homomorphically encrypted data can support real-time learning in personalized cancer medicine. BMC Med Inform Decis Mak 2019;19(1):255.
8. Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. BMC Med Inform Decis Mak 2019;19(1):44.
9. Schüssler-Fiorenza Rose SM, Contrepois K, Moneghetti KJ, Zhou W, Mishra T, Mataraso S, et al. A longitudinal big data approach for precision health. Nat Med 2019;25(5):792–804.
10. Fang G, Annis IE, Elston-Lafata J, Cykert S. Applying machine learning to predict real-world individual treatment effects: insights from a virtual patient cohort. J Am Med Inform Assoc 2019;26 (10):977–88.
11. Zhang Y, Cai T, Yu S, Cho K, Hong C, Sun J, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc 2019;14(12):3426–44.
12. Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. Int J Med Inf 2019;129:13–9.
13. Miller HN, Gleason KT, Juraschek SP, Plante TB, Lewis-Land C, Woods B, et al. Electronic medical record-based cohort selection and direct-to-patient, targeted recruitment: early efficacy and lessons learned. J Am Med Inform Assoc 2019;26(11):1209–17.
14. Claerhout B, Kalra D, Mueller C, Singh G, Ammour N, Meloni L, et al. Federated electronic health records research technology to support clinical trial protocol optimization: Evidence from EHR4CR and the InSite platform. J Biomed Inform 2019;90:103090.
15. Carrigan G, Whipple S, Capra WB, Taylor MD, Brown JS, Lu M, et al. Using Electronic Health Records to Derive Control Arms for Early Phase Single-Arm Lung Cancer Trials: Proof-of-Concept in Randomized Controlled Trials. Clin Pharmacol Ther 2020;107(2):369–77.
16. Beier K, Schweda M, Schicktanz S. Taking patient involvement seriously: a critical ethical analysis of participatory approaches in data-intensive medical research. BMC Med Inform Decis Mak 2019;19(1):90.

Correspondence to:
Christel Daniel, MD, PhD
Data and Digital Innovation Department, Information Systems Direction – 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

Appendix: Summary of Best Papers Selected for the 2020 Edition of the IMIA Yearbook, CRI Section

Paddock S, Abedtash H, Zummo J, Thomas S

Proof-of-concept study: Homomorphically encrypted data can support real-time learning in personalized cancer medicine

BMC Med Inform Decis Mak 2019 Dec 4;19(1):255

Data protection, and in particular complying with the GDPR, when reusing real world data for research is recognised to be challenging. This can be an important barrier to scaling up personalised medicine research, when it may be difficult or impossible to robustly anonymise very fine-grained health and treatment history and biological profiles. This study applied the technique of Homomorphic Encryption (HE) to the problem, which allows the source data to remain encrypted when a research query is executed. The only knowledge needed in advance would be the structure of the database and its semantics such as the terminology systems used. Analysis using HE means that personal data does not need to be disclosed to any person or computational process during query execution. In this proof of concept study, a previously published HE technique was used to encrypt a database containing 5000 simulated patient records modelled on personalized medicine treatment scenarios. The authors sought to detect outlying (and therefore rare) drug responses from the data. Exceptional treatment response queries were performed directly on the encrypted data via HE, using conventionally available computing facilities. They found that the queries were able to return correct responses, permitting their sample research questions about exceptional drug response to be answered. Although this method is more time consuming to perform the queries, the authors argue that this computational time is relatively small within the total time frame of a personalised medicine study. This research was included as a best paper because it demonstrates a valuable method for more data sharing and distributed data querying in cases where the fine granularity of the clinical/molecular data does not permit robust anonymisation.

Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, Reich CG, Duke J, Madigan D, Hripcsak G, Ryan PB

Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis

Lancet 2019 Nov 16;394(10211):1816-26

At present, clinical guidelines for the treatment of hypertension recommend several different classes of drug in patients without a comorbidity or significant risk factor. Thiazide or thiazide-like diuretics, angiotensin-converting enzyme (ACE) inhibitors, angiotensin receptor blockers, dihydropyridine calcium channel blockers, and non-dihydropyridine calcium channel blockers are all candidates for consideration, with no prioritisation amongst this list. Given that the difference in treatment effects between them may be very small, a classical Randomised Controlled Trial would not be feasible to identify if one of these candidates is more effective. This paper reports the results of a large-scale real world data study to identify an effectiveness difference and also to demonstrate the potential value of a federated health data network for big data observational research. The authors undertook a distributed analysis of 4.9 million patient records in retrospective cohorts. These patients were from nine claims and EHR databases in the US, Japan, South Korea and Germany, within the LEGEND-HTN study which is comparing antihypertensive drug treatments. The analysis was undertaken through the Observational Health Data Science and Informatics (OHDSI) distributed data network. This platform mapped all of the data sources to the Observational Medical Outcomes Partnership (OMOP) common data model, which permitted uniform federated (distributed) analysis queries to be executed. Fifty-five health outcomes were studied. Thiazide-like diuretics had better effectiveness than ACE inhibitors to reduce the incidence of acute myocardial infarction, hospitalisation for heart failure, and stroke risk. This research was included as a best paper because it reports the largest scale study on this topic, which was only possible because of the federated database network, coupled with a robust study methodology. It demonstrates the evidence-generation value of federated health data networks.

Yu Y, Ruddy KJ, Hong N, Tsuji S, Wen A, Shah ND, Jiang G

ADEpedia-on-OHDSI: A next generation pharmacovigilance signal detection platform using the OHDSI common data model

J Biomed Inform 2019 Mar;91:103119

Underreporting of adverse drug events (ADEs) is a key challenge in drug safety surveillance. Although a valuable resource for pharmacovigilance, the US Food and Drug Administration (FDA)'s Adverse Event Reporting System (FAERS) only capture adverse drug reactions (ADRs) spontaneously reported by healthcare professionals, patients, and pharmaceutical manufacturers. Longitudinal observational databases like Electronic Health Records (EHRs) and transactional claims can be used as additional data sources for pharmacovigilance to address gaps in coverage and increase population heterogeneity. In order to integrate the data from those two types of sources - spontaneous reporting system (SRS) and EHRs – with different data models and vocabularies, Yu *et al.,* considered the use of the OMOP common data model (CDM). This model is not only increasingly adopted by data research networks leveraging EHRs data but has also been intensively used to identify and assess associations between medical interventions and health-related outcomes in many pharmacovigilance and pharmacoepidemiology studies. The authors converted into the OMOP format the last version of the FEARS data base (including 4,619,362 adverse event cases reported between 2012 and 2017). A dedicated tool has been developed to extract, transform, and load (ETL) the FEARS data into the OMOP data base (version 5). An important part of the ETL process was dedicated to terminology mappings of the drug names to RxNorm

and of the adverse events, indications, and outcomes to SNOMED CT. The structure mapping between FEARS tables and OMOP CDM required multiple rounds of discussion involving two experts with medical informatics background. The evaluation of the work was two-fold. The authors first validated the mappings and the conversion process and conducted a replication study in order to evaluate the impact of the conversion and information loss on signal detection.

This paper was selected as a best paper because it demonstrates with a robust methodology not only the feasibility of converting a SRS data base into the OMOP format but also the accuracy of the resulting framework called ADEpedia-on-OHDSI and its capability to improve signal detection through standardization. Furthermore, this work paves the way for seamless integration of SRS with EHRs or other RWD enabling better signal detection and further discovery about adverse events such as causes, confounders, or possible corrective actions.