Clinical Research Informatics

Christel Daniel[1,2], Xavier Tannier[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 Université, Université Sorbonne Paris Nord, INSERM, 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 2021.

Method: Using PubMed, we did a bibliographic search using a combination of MeSH descriptors and free-text terms on CRI, followed by a double-blind review in order to select a list of candidate best papers to be peer-reviewed by external reviewers. After peer-review ranking, three section editors met for a consensus meeting and the editorial team was organized to finally conclude on the selected three best papers.

Results: Among the 1,096 papers (published in 2021) returned by the search and in the scope of the various areas of CRI, the full review process selected three best papers. The first best paper describes an operational and scalable framework for generating EHR datasets based on a detailed clinical model with an application in the domain of the COVID-19 pandemics. The authors of the second best paper present a secure and scalable platform for the preprocessing of biomedical data for deep data-driven health management applied for the detection of pre-symptomatic COVID-19 cases and for biological characterization of insulin-resistance heterogeneity. The third best paper provides a contribution to the integration of care and research activities with the REDCap Clinical Data and Interoperability Services (CDIS) module improving the accuracy and efficiency of data collection.

Conclusions: The COVID-19 pandemic is still significantly stimulating research efforts in the CRI field to improve the process deeply and widely for conducting real-world studies as well as for optimizing clinical trials, the duration and cost of which are constantly increasing. The current health crisis highlights the need for healthcare institutions to continue the development and deployment of Big Data spaces, to strengthen their expertise in data science and to implement efficient data quality evaluation and improvement programs.

Keywords

Clinical Trials as Topic; Observational Studies as Topic; Real-world Data; Real-world Evidence Generation; Phenotyping

Yearb Med Inform 2022:161-6
http://dx.doi.org/10.1055/s-0042-1742530

1 Introduction

For the 2021 International Medical Informatics Association (IMIA) Yearbook, the goal of the Clinical Research Informatics (CRI) section is to provide an overview of research trends from 2021 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 life-cycle of clinical trials. The CRI community has also addressed the important challenges of addressing bias and equity and contributed to "Inclusive Digital Health" –this year's special theme for the IMIA Yearbook.

2 Paper Selection Method

A comprehensive review of articles published in 2021 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 8, 2022 for papers published in 2021 and considering the electronic publication date. Among an original set of 1,096 references, 1,048 papers were selected as being in the scope of CRI and their scientific quality was blindly rated as low, medium, or high by the three section editors based on papers' title and abstract. Seventy-four references classified as high quality contributions to the field by at least two of the three section editors were considered and classified into the following eleven areas of the CRI domain in order of the number of matching papers (multiple classification choices were permitted): reuse of Electronic Healthcare Records (EHRs); Learning Healthcare System (LHS) data; Big data management, data integration, semantic interoperability and data quality assessment; Data science (data/text mining, Artificial Intelligence (IA), Machine

Learning (ML)); Security and data privacy; and feasibility studies, patient recruitment, improved user experiences of CRI systems and governance (ethical, regulatory, societal, policy issues, stakeholder participation, research networks, team science).
The 74 references were reviewed jointly by the 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 2 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2022 in the section 'Clinical Research Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Clinical Research Informatics
▪ Bahmani A, Alavi A, Buergel T, Upadhyayula S, Wang Q, Ananthakrishnan SK, Alavi A, Celis D, Gillespie D, Young G, Xing Z, Nguyen MHH, Haque A, Mathur A, Payne J, Mazaheri G, Li JK, Kotipalli P, Liao L, Bhasin R, Cha K, Rolnik B, Celli A, Dagan-Rosenfeld O, Higgs E, Zhou W, Berry CL, Van Winkle KG, Contrepois K, Ray U, Bettinger K, Datta S, Li X, Snyder MP. A scalable, secure, and interoperable platform for deep data-driven health management. Nat Commun 2021 Oct 1;12(1):5757.
▪ Cheng AC, Duda SN, Taylor R, Delacqua F, Lewis AA, Bosler T, Johnson KB, Harris PA. REDCap on FHIR: Clinical Data Interoperability Services. J Biomed Inform 2021 Sep;121:103871.
▪ Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, Terriza-Torres AI, López-Jiménez EA, Calvo-Boyero F, Jiménez-Cerezo MJ, Blanco-Martínez AJ, Roig-Domínguez G, Cruz-Bermúdez JL, Bernal-Sobrino JL, Serrano-Balazote P, Muñoz-Carrero A. Obtaining EHR derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. J Biomed Inform 2021 Mar;115:103697.

3 Conclusions and Outlook

The 15 candidate best papers for 2021 illustrate recent efforts and trends in different CRI areas such as real-world evidence generation; Data integration, semantic interoperability and data quality assessment; security, confidentiality and data privacy; data/text mining, Artificial Intelligence (AI) and Machine Learning (ML); feasibility studies, patient recruitment, data management and CRI systems; and ethical, legal, social, policy issues and solutions.

3.1 Real-world Evidence Generation, Electronic Phenotyping

The increasing scale and scope of biomedical data is not only generating enormous opportunities for improving health outcomes but also raises new challenges ranging from data acquisition and storage to data analysis and utilization. The first best paper of the CRI section, from Bahmani et al. describes the Personal Health Dashboard (PHD), an open-source framework using state of the art security and scalability technologies that can be deployed to big data projects [1]. The PHD was prototyped to enable the collection and visualization of diverse biomedical data types (wearable, clinical, omics) at a personal level, the investigation of insulin resistance and the detection of pre-symptomatic COVID-19. The large amount of real-world health data increasingly accessible is an opportunity for conducting nonrandomized controlled studies (NRCS). Various recently published meta-epidemiological studies comparing treatment effect estimates from nonrandomized controlled studies (NRCS) and Randomized controlled trials (RCTs) showed mixed results. Mathes *et al.*, conducted the first meta-epidemiological study trying to get deeper insights into the actual causes of disagreement in treatment effect estimates between NRCS based on real-world data and RCTs [2]. They identified many potential causes of disagreement in treatment effect estimates between NRCS and RCTs but only few of them would probably have resulted in a different conclusion regarding the harm/benefit of the intervention in practice.

3.2 Data Integration, Semantic Interoperability and Data Quality Assessment

Making real-world data interoperable and reusable is a major challenge in the CRI field to unlock the potential of the data to support research and innovation. In the second best paper of the section, Pedrera-Jimenez *et al.* describe an operational and scalable framework for rapidly and efficiently obtaining EHR-derived data for secondary use in COVID-19, capable of adapting to changes in data specifications and ensuring acceptable data quality [3]. The framework based on detailed clinical models (parts 1 and 2 of the ISO 13606 standard) semantically linked to standards such as SNOMED-CT and LOINC and was applied to extract and transform data from the EHR. The data set built conformed to the ISARIC-WHO COVID-19 case report form was built for a population of 4,489 COVID-19 patients with an acceptable data completeness and without requiring manual data collection. Methodologies and tools to assess data quality are still an active research area of the field. Verma *et al.*, describe the data quality assessment process of the GEMINI database containing 245,559 patient admissions at seven hospitals in Ontario, Canada from 2010 to 2017 [4]. Nearly 90% of data quality issues were related to data extraction and transfer from hospitals rather than processing at the central site. Correcting these issues required strong support from staff and leadership at participating hospitals. The authors highlight the importance of an iterative data quality assessment methodology that sequentially combines computational most effective at detecting systematic errors (e.g., a large chunk of missing data or all dates/times shifted by a fixed amount) and manual techniques to address non-systematic errors. The authors reduced the number of data points to be manually validated by focusing on key data variables through a "fit-for-purpose" approach. To reinforce the trust in real-world data, and especially in the databases standardized using the OMOP Common Data Model, Blacketer *et al.*, developed a data quality dashboard (DQD), an open-source R package reporting potential quality issues through the

systematic execution and summarization of over 3,300 configurable data quality checks [5]. Adopting the harmonized data quality assessment (DQA) framework from Kahn *et al.*, Kapsner *et al.*, developed a DQA tool linked to common data element definitions stored in a metadata repository (MDR) and deployed it within the MIRACUM consortium [6].

3.3 Data/Text Mining, Artificial Intelligence and Machine Learning

Machine Learning (ML) and Natural Language Processing (NLP) tools are now an integral part of large-scale efforts to leverage real-world data for different purposes in addition to statistical methods [7,8]. The Personal Health Dashboard mentioned earlier incorporates prediction of phenotypic traits such as insulin resistance through a logistic regression algorithm [1]. Dhayne *et al.*, propose a representation learning NLP pipeline for finding links between medical reports and clinical reports [9]. They combine term extraction from both the reports and the trial dataset, and ontological reasoning techniques, in order to build a vector representation of the documents and thus to match a patient to a clinical trial. While classical statistics have an arsenal of methods to validate hypotheses and estimate the necessary size of the study population, machine learning approaches fall short on these aspects. Liu *et al.*, allow progress to be made in this area [10]. They propose sample size formulae based on pre-specified lower bounds for the objective metrics, as well as a stepwise strategy for iterative algorithm development/validation cycles.

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

There are now many tools and products supporting the reuse of EHR data to address insufficient patient recruitment in clinical trials. It remains challenging to match EHR data to eligibility criteria, whose definition process remains opaque, unscalable and insufficiently inclusive. Liu *et al.*, developed the clinical trial knowledge base (CTKB), a novel comprehensive and regularly updated knowledge base of discrete eligibility criteria concepts with the potential to enable knowledge engineering for clinical trial cohort definition, clinical trial population representativeness assessment, electronical phenotyping to support clinical trial recruitment [11]. Interestingly, Rogers *et al.*, propose a generalizability assessment method to compare between trial participants and potentially eligible patients using electronic health record data [12]. The third best paper of the CRI section, from Cheng *et al.*, addresses the time-consuming and error prone re-capture of patient histories from EHR to electronic data capture (EDC) systems [13]. The Clinical Data and Interoperability Services (CDIS) are based on generalizable modules for real-time data exchange between vendor EHR systems and an electronic data capture system (REDCap) developed in a scalable manner to facilitate clinical and translational research. By leveraging HL7 FHIR standards, these new REDCap modules, enabling data extraction for traditional CRF-based studies, registries, and data marts, have been successfully deployed and used at Vanderbilt University Medical Center and disseminated to many REDCap institutions.

3.5 Ethical, Legal, Social, Policy Issues and Solutions, Stakeholder Participation, Research Networks

There is growing recognition of the importance of the ethical development of health informatics research and innovation. The area gaining maximal attention at present is the development of artificial intelligence, with many countries publishing ethical principles, guidelines, or more formal legislation (such as the forthcoming European AI Act 1). This is the area covered by our literature survey paper in this chapter, "A Literature Review on Ethics for AI in Biomedical Research and Biobanking". The authors found that the ethical principles were similar across biomedical research and bio-banking, and specifically looked for publications that provided practical guidance on how to follow ethical principles, since this is what most of the research community need in order to be confident themselves and to give confidence to others that they are correctly following ethical principles. They found "only a few publications dedicated to helping practitioners with implementation of these high-level principles in practice", which they have summarized. Research Ethics Committees have also come under scrutiny in a recent publication by Ferretti et al, with a particular critique of their capability to adequately assess big data research studies [14]. This is not the scope for which they were originally constituted, and the authors have listed a number of areas of weakness, for example inconsistent decision making for multi-site data studies, a lack of expertise in appraising proposals for anonymisation, the appropriate wording and terms of data related consent, and a challenge in knowing how to formulate and evaluate data related risks. The authors call for reform and present a set of structural reforms and areas of guidance that ethics committee members should have access to.

On the topic of research networks, a detailed overview of the establishment and growth of PCORnet (the US Patient-Centered Clinical Research Network) was recently published by Forrest *et al.* [15]. This publication explains the PCORnet mission, to establish a "network-of-networks that engages patients, caregivers, clinicians, health system leaders, payers, and researchers in the design, conduct, and advancement of patient-centered outcomes research", how it is constituted and the way in which it engages with relevant stakeholders. The authors profile the 80 million patients, reflecting population diversity, who are now represented within this distributed (multi-site) network, with data that is now ready for outcomes research studies.

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 for this IMIA Yearbook.

[1] Proposal for a Regulation of the European Parliament and of The Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

References

1. Bahmani A, Alavi A, Buergel T, Upadhyayula S, Wang Q, Ananthakrishnan SK, et al. A scalable, secure, and interoperable platform for deep data-driven health management. Nat Commun 2021 Oct 1;12(1):5757.
2. Mathes T, Rombey T, Kuss O, Pieper D. No inexplicable disagreements between real-world data-based nonrandomized controlled studies and randomized controlled trials were found. J Clin Epidemiol 2021 May;133:1-13.
3. Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, Terriza-Torres AI, López-Jiménez EA, Calvo-Boyero F, et al. Obtaining EHR derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. J Biomed Inform 2021 Mar;115:103697.
4. Verma AA, Pasricha SV, Jung HY, Kushnir V, Mak DYF, Koppula R, et al. Assessing the quality of clinical and administrative data extracted from hospitals: the General Medicine Inpatient Initiative (GEMINI) experience. J Am Med Inform Assoc 2021 Mar 1;28(3):578-87.
5. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc 2021 Sep 18;28(10):2251-7.
6. Kapsner LA, Mang JM, Mate S, Seuchter SA, Vengadeswaran A, Bathelt F, et al. Linking a Consortium-Wide Data Quality Assessment Tool with the MIRACUM Metadata Repository. Appl Clin Inform 2021 Aug;12(4):826-35.
7. Chen WC, Li H, Wang C, Lu N, Song C, Tiwari R, et al. Evaluation of diagnostic tests for low prevalence diseases: a statistical approach for leveraging real-world data to accelerate the study. J Biopharm Stat 2021 May 4;31(3):375-90.
8. Schulz WL, Young HP, Coppi A, Mortazavi BJ, Lin Z, Jean RA, et al. Temporal relationship of computed and structured diagnoses in electronic health record data. BMC Med Inform Decis Mak 2021 Feb 17;21(1):61.
9. Dhayne H, Kilany R, Haque R, Taher Y. EMR2vec: Bridging the gap between patient data and clinical trial. Comput Ind Eng 2021 Jun;156:107236.
10. Liu L, Bustamante R, Earles A, Demb J, Messer K, Gupta S. A strategy for validation of variables derived from large-scale electronic health record data. J Biomed Inform 2021 Sep;121:103879.
11. Liu H, Chi Y, Butler A, Sun Y, Weng C. A knowledge base of clinical trial eligibility criteria. J Biomed Inform 2021 May;117:103771.
12. Rogers JR, Hripcsak G, Cheung YK, Weng C. Clinical comparison between trial participants and potentially eligible patients using electronic health record data: A generalizability assessment method. J Biomed Inform 2021 Jul;119:103822.
13. Cheng AC, Duda SN, Taylor R, Delacqua F, Lewis AA, Bosler T, et al. REDCap on FHIR: Clinical Data Interoperability Services. J Biomed Inform 2021 Sep;121:103871.
14. Forrest CB, McTigue KM, Hernandez AF, Cohen LW, Cruz H, Haynes K, et al. PCORnet® 2020: current state, accomplishments, and future directions. J Clin Epidemiol 2021 Jan;129:60-7.
15. Ferretti A, Ienca M, Sheehan M, Blasimme A, Dove ES, Farsides B, et al. Ethics review of big data research: What should stay and what should be reformed? BMC Med Ethics 2021 Apr 30;22(1):51.

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