Published online: 2019-08-16

Clinical Research Informatics: Contributions from 2018

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

[1] AP-HP Information Systems Direction, 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 2018.

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 of the editorial team was organized to conclude on the selection of best papers.

Results: Among the 1,469 retrieved papers published in 2018 in the various areas of CRI, the full review process selected four best papers. The first best paper describes a simple algorithm detecting co-morbidities in Electronic Healthcare Records (EHRs) using a clinical data warehouse and a knowledge base. The authors of the second best paper present a federated algorithm for predicting heart failure hospital admissions based on patients' medical history described in their distributed EHRs. The third best paper reports the evaluation of an open source, interoperable, and scalable data quality assessment tool measuring completeness of data items, which can be run on different architectures (EHRs and Clinical Data Warehouses (CDWs) based on PCORnet or OMOP data models). The fourth best paper reports a data quality program conducted across 37 hospitals addressing data quality issues through the whole data life cycle from patient to researcher.

Conclusions: Research efforts in the CRI field currently focus on consolidating promises of early Distributed Research Networks aimed at maximizing the potential of large-scale, harmonized data from diverse, quickly developing digital sources. Data quality assessment methods and tools as well as privacy-enhancing techniques are major concerns. It is also notable that, following examples in the US and Asia, ambitious regional or national plans in Europe are launched that aim at developing big data and new artificial intelligence technologies to contribute to the understanding of health and diseases in whole populations and whole health systems, and returning actionable feedback loops to improve existing models of research and care. The use of "real-world" data is continuously increasing but the ultimate role of this data in clinical research remains to be determined.

Keywords

IMIA Yearbook, clinical research informatics, biomedical research, clinical trials as topic, observational studies as topic, phenotype

Yearb Med Inform 2019:203-7
http://dx.doi.org/10.1055/s-0039-1677921

Introduction

Within the 2018 International Medical Informatics Association (IMIA) Yearbook, the Clinical Research Informatics (CRI) section aims at providing an overview of research trends from 2018 publications that demonstrate the progress in multifaceted aspects of medical informatics supporting the life-cycle of clinical trials as well as the always growing use of "real-world" data. New methods, tools, and CRI systems have been developed in order to collect, integrate, and mine healthcare data for better care. The CRI community has especially addressed the important challenges of evaluating the impact of "new artificial intelligence technologies", this year's special theme of the IMIA Yearbook.

About the Paper Selection

A comprehensive review of articles published in 2018 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, Phenotype, 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, 2019 for papers published in 2018, considering the electronic publication date. Among an original set of 1,468 references, 1,019 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-four references classified as medium or high quality contributions to the field by at least one of the section editors were classified into the following eleven dimensions/sub areas of the CRI domain: observational studies, reuse of electronic health record (EHR) data, data integration

and semantic interoperability, feasibility studies, patient recruitment, data management and CRI systems, data/text mining and algorithms, data quality assessment or validation, security and confidentiality, ethical, legal, social, policy issues and solutions, stakeholder participation, communicating study results. The 84 references were reviewed jointly by the two section editors to select a consensual list of 14 candidate best papers representative of all CRI categories. Following the IMIA Yearbook process, these 14 papers were peer-reviewed by the IMIA Yearbook 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 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 2019 in the section 'Clinical Research Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section
Clinical Research Informatics
▪ Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated Electronic Health Records. Int J Med Inform 2018 Apr;112:59-67.
▪ Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed 2018 Nov 9.
▪ Estiri H, Stephens KA, Klann JG, Murphy SN. Exploring completeness in clinical data research networks with DQe-c. J Am Med Inform Assoc 2018 Jan 1;25(1):17-24.
▪ Sylvestre E, Bouzillé G, Chazard E, His-Mahier C, Riou C, Cuggia M. Combining information from a clinical data warehouse and a pharmaceutical database to generate a framework to detect comorbidities in electronic health records. BMC Med Inform Decis Mak 2018 Jan 24;18(1):9.

Outlook

The 14 candidate best papers for 2018 illustrate recent efforts towards data-driven research and innovation and exemplify trends in CRI sub-domains such as data/text mining, artificial intelligence, data integration and semantic interoperability, data management and CRI systems, data quality and reproducibility in biomedical research, security, initiatives for scaling up real world data. In addition to these research papers, a useful overview of the challenges and approaches to scaling up research using large-scale health data resources was published by Hemingway *et al*. [1].

Data/Text Mining and New Technologies from Artificial Intelligence

The proliferation of diverse health data sources has made feasible the analysis of "real-world" data to generate evidence for healthcare professional decision-making. For example, Ledieu *et al.* demonstrates that smart representation of heterogeneous data integrated within Clinical Data Warehouses (CDWs) improves care givers' experience [2]. One of the **best papers** is a paper from Sylvestre *et al.,* [3]. The authors propose an algorithm to detect comorbidities in electronic health records (EHRs). It combines structured data such as drug prescriptions and laboratory results with indications for each drug provided by a pharmaceutical database. Comorbidity diagnoses were suggested for 68.4% of the 4,312 patients of the test data set and confirmed in 20.3% of reviewed cases. Important health information in hospital CDWs is hidden in unstructured data. Garcelon *et al.*, [4] have combined two information extraction methods to detect phenotypes for patients with rare diseases. The document-oriented CDW PaDaWaN has been extended by Dietrich *et al.*, [5] with an ad hoc dynamic, interactive, and adjustable information extraction service that allows users to query text data in a manner similar to the one used to query structured data. This works on the fly, at runtime, to recognize negation and context, and can compute the frequencies for Boolean and numeric values with high recall and precision.

One method for data protection of federated (virtual) databases is by avoiding granular data exchange. Another **best paper** is a paper authored by Brisimi *et al.,* [6] which describes a computationally efficient and privacy-aware solution for large-scale machine learning problems running on distributed data. The iterative cluster Primal Dual Splitting (cPDS) algorithm, developed for solving the large-scale sparse Support Vector Machine (sSVM) problem in a decentralized fashion allows the data holders to collaborate, while keeping every participant's data private. The distributed learning scheme cPDS, evaluated on the problem of predicting hospitalizations due to heart diseases, converges faster than centralized methods and achieves similar prediction accuracy.

Data Integration and Semantic Interoperability

Data heterogeneity is one of the critical problems in sharing or linking, reusing, and analysing datasets. Fast Healthcare Interoperability Resources (FHIR) is the new HL7 interoperability standard. Substitutable Modular third-party Applications (SMART) defines the SMART-on-FHIR specification for how applications interface with EHRs through FHIR. Paris *et al.,* [7] extended i2b2 to search remotely into one or multiple SMART-on-FHIR Application Programming Interfaces (APIs). This opens i2b2 to new data types and improves security and interoperability management in the context of scalable solutions for cross-border and cross-domain networking of data.

Data Management and CRI Systems

Devine *et al.,* [8] present an evaluation of data management at the hospitals of the Washington State's Surgical Care Outcomes and Assessment Program (SCOAP) network engaged in the Comparative Effectiveness Research and Translation Network (CERTAIN). It aims at reusing EHRs for quality improvement and research. The authors compared a manual and an automated abstraction processes based on a centralized federated data model in four SCOAP hospitals. Six to 15 percent of data elements were automatically abstracted with more than 90% of consistency.

Data Quality and Reproducibility in Biomedical Research

Although a major concern in distributed research networks (DRNs), data quality (DQ) assessment of hospital information systems is largely unpublished. The US National Patient-Centered Clinical Research Network (PCORnet®) is one of the first DRNs to incorporate EHR data from multiple domains on a national scale. Quals *et al.,* [9] describe the data curation process of the PCORnet's Coordinating Center for evaluating foundational DQ and assessing fitness-for-use across a broad research portfolio.

Looten *et al,* [10] leveraged the European Hospital Georges Pompidou CDW and tracked the evolution of 192 biological parameters over 17 years (445,000+ patients, 131 million laboratory test results). The authors developed computational and statistical methods to identify different evolution profiles and formulated recommendations to enable safe use and sharing of biological data collection to limit the impact of data evolution in retrospective and federated studies.

The paper from Daniel *et al.,* [11] selected as a **best paper**, presents a DQ program at AP-HP to increase the reproducibility of analyses running on the CDW aggregating EHR data from 37 hospitals. Two DQ campaigns were conducted in patient identification (PI) and healthcare services (HS). The results of the semi-automated DQ profiling in the PI data set (8.8 M patients) and the HS data set (13,099 consultations, 2,122 care units) are presented with improvement campaigns that have already resulted in significant DQ improvement (11).

The paper from Estiri *et al*., [12], also selected as a **best paper**, presents DQe-c, an open source, interoperable, and scalable data quality assessment tool for evaluation and visualization of completeness and conformance in EHR data repositories based on either the PCORnet® or OMOP Common Data Model. DQe-c was validated on 200 000 patient records randomly selected from the Research Patient Data Registry at Partners HealthCare. The web-based DQ reports include descriptive graphics and tables that are tailored to EHR DQ assessment but could be extended to the other steps of the data quality life-cycle.

Security

Linking record-level data between repositories often utilises a pseudonym (a linkage key), for which privacy preserving linkage is an important approach to enable compliance with the EU General Data Protection Regulation (GDPR). A paper in 2018 applies the secure Multi-Party Computation (MPC), a well-known technique for Privacy-Preserving Data Mining, to three pilot data mining scenarios: location tracking within a hospital; joint data analysis across multiple care providers; mining a mixture of data sources [13]. MPC is proposed as a scalable method for linked data mining in a GDPR compliant way.

Initiatives for Scaling up Real World Data

Several European countries, alongside others globally, are investing in national infrastructures and competencies to integrate EHR data at scale to enable big data research. The two newest programmes to be launched are in Germany [14] and France. They have been designed quite differently, and the Survey Paper in this section provides an in depth analysis and comparison of both initiatives [15]. There are valuable opportunities for both programmes to learn from each other.

Acknowledgement

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

References

1. Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. Eur Heart J 21 avr 2018;39(16):1481-95.
2. Ledieu T, Bouzillé G, Thiessard F, Berquet K, Van Hille P, Renault E, et al. Timeline representation of clinical data: usability and added value for pharmacovigilance. BMC Med Inform Decis Mak 2018;18(1):86.
3. Sylvestre E, Bouzillé G, Chazard E, His-Mahier C, Riou C, Cuggia M. Combining information from a clinical data warehouse and a pharmaceutical database to generate a framework to detect comorbidities in electronic health records. BMC Med Inform Decis Mak 24 2018;18(1):9.
4. Garcelon N, Neuraz A, Salomon R, Bahi-Buisson N, Amiel J, Picard C, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet J Rare Di. 31 2018;13(1):85.
5. Dietrich G, Krebs J, Fette G, Ertl M, Kaspar M, Störk S, et al. Ad Hoc Information Extraction for Clinical Data Warehouses. Methods Inf Med 2018;57(1):e22-9.
6. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated Electronic Health Records. Int J Med Inf 2018;112:59-67.
7. Paris N, Mendis M, Daniel C, Murphy S, Tannier X, Zweigenbaum P. i2b2 implemented over SMART-on-FHIR. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci 2018;2017:369-78.
8. Devine EB, Van Eaton E, Zadworny ME, Symons R, Devlin A, Yanez D, et al. Automating Electronic Clinical Data Capture for Quality Improvement and Research: The CERTAIN Validation Project of Real World Evidence. EGEMS Wash DC 22 mai 2018;6(1):8.
9. Qualls LG, Phillips TA, Hammill BG, Topping J, Louzao DM, Brown JS, et al. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). eGEMs [Internet]. [cité 9 déc 2018];6(1). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5983028/
10. Looten V, Kong Win Chang L, Neuraz A, Landau-Loriot M-A, Vedie B, Paul J-L, et al. What can millions of laboratory test results tell us about the temporal aspect of data quality? Study of data spanning 17 years in a clinical data warehouse. Comput Methods Programs Biomed 2018 Dec 29.
11. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed. 2018 Nov 9.
12. Estiri H, Stephens KA, Klann JG, Murphy SN. Exploring completeness in clinical data research networks with DQe-c. J Am Med Inform Assoc 2018;25(1):17-24.
13. Veeningen M, Chatterjea S, Horváth AZ, Spindler G, Boersma E, van der Spek P, et al. Enabling Analytics on Sensitive Medical Data with Secure Multi-Party Computation. Stud Health Technol Inform 2018;247:76-80.
14. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018;57(S 01):e50-6.
15. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics Initiative: Two National Projects to Promote Data Sharing in Healthcare. Yearb Med Inform 2019:190-202.

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 2019 IMIA Yearbook, Section Clinical Research Informatics

Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W

Federated learning of predictive models from federated Electronic Health Records

Int J Med Inform 2018 Apr;112:59-67

Centralized machine learning methods are typically used to train predictive models from data that are aggregated into large central repositories. This paper describes an alternative machine learning method applicable to massive data residing in different locations and owned by different entities that could not be aggregated into a single repository due to technical and/or privacy concerns. Brisimi *et al.* developed a new federated algorithm - the cluster Primal Dual Splitting (cPDS) algorithm - for solving the large-scale sparse Support Vector Machine (sSVM) problem in a decentralized fashion. They applied this new algorithm to a dataset of de-identified Electronic Healthcare Records (EHRs) from the Boston Medical Center for predicting heart failure hospital admissions based on patients' medical history described in their distributed EHRs. The federated optimization scheme cPDS enables multiple data holders to collaborate and converge to a common predictive model, without explicitly exchanging raw data. This distributed algorithm accurately differentiates between patients that are likely or unlikely to be hospitalized within a target year. With a prediction accuracy of 0,7806 measured by the Area Under the Receiver Operating Characteristic Curve (AUC), the cPDS performs better than the other methods used to solve the sSVM problem. The authors demonstrate that cPDS converges faster than both centralized methods and alternative distributed algorithm. Important features that are predictive of future hospitalizations have been discovered by the algorithm, such as age, diagnosis of heart failure in the year before the target year, admission due to heart failure or other circulatory system diagnoses one year before the target year, thus providing a way to interpret the classification results and inform prevention efforts. At a time of increasing preference for distributed querying to avoid centralizing data, this paper makes a great contribution by providing a validated approach to distributed machine learning for such architectures.

Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N

Initializing a hospital-wide data quality program. The AP-HP experience

Comput Methods Programs Biomed 2018 Nov 9

There is increasing recognition of the importance of assessing and improving the data quality (DQ) of hospital EHRs, not only for clinical care purposes but to enable robust clinical research inferences from the data. This is of of the two 2018 best papers selected on DQ for this section. Unlike many publications that focus only on assessment, this paper has been selected because it tackles the challenge holistically examining how DQ can be improved, and undertook the research across 37 hospitals in the Paris region (the Assistance Publique – Hôpitaux de Paris, AP-HP). Daniel *et al,.* designed and conducted DQ campaigns consisting of five phases: defining the scope, measuring, analyzing, improving, and controlling DQ. They applied this in two domains - patient identification and healthcare services. Through EHR data profiling across the AP-HP network, comprising a repository of 8.8 million patients, the authors identified 11 data quality issues. These were categorized into completeness, conformance, and plausibility DQ issues. The root causes of these issues were found to be errors from data originators, ETL issues, or limitations of the source EHR data, and these insights informed a DQ improvement campaign. The improvement strategies targeted staff communication and teaching (leaflets, videos, feedback), the engagement of patient registration staff and health professionals (DQ campaigns, updating specialty vocabulary tables), patient engagement (in rechecking their information), and information system improvements such as computerised DQ checks and fixing record merger errors. These action plans, though only partially implemented at the time of publication, resulted in significant improvement of DQ measures. This research was included as a best paper because it provides insights into the actual data quality observed across 37 hospitals, linked to the kinds of campaign actions that can be implemented: the research goes beyond assessment to improvement.

Estiri H, Stephens KA, Klann JG, Murphy SN

Exploring completeness in clinical data research networks with DQe-c

J Am Med Inform Assoc 2018 Jan 1;25(1):17-24

A new generation of clinical research platforms offers the capability to reuse on a large (multi-site, federated) scale routinely collected hospital EHR data for clinical research, such as clinical trials and big data mining. Since data quality (DQ) imperatives to support continuity of care and to support reuse for research are quite different, DQ poses important concerns for secondary use of EHR data and remains a challenge for research data networks using non-scalable ad hoc solutions. Estiri's paper is one of the two 2018 best papers selected on DQ for this section. The authors developed an open source, interoperable, and scalable DQ assessment tool able to measure the completeness and conformance of data items within an EHR or CDW data model. They describe the iterative implementation of a web-based tool - DQe-c - across different institutions focusing on interoperability and scalability to large databases. The DQe-c has been evaluated on a sample dataset of 200,000 randomly selected patient records with an encounter since January 1, 2010, extracted from the Research Patient Data Registry at Partners HealthCare. The web-based report produced by DQe-c is organized into four sections: load and test details (list of the tables of the EHR or CDW's common data model (CDM) and table-level size and completeness), completeness test (missing data for each column of the tables), data model conformance test (rate of orphan records based on the CDM), and fitness for purpose test (missingness in key clinical indicators such as ethnicity

data or blood pressure for example). This best paper contributes to the body of open algorithms and tools to permit comparable DQ assessments which can be run on different architectures (EHRs, PCORnet, OMOP) paving the way to systematic evaluation of DQ across distributed networks.

Sylvestre E, Bouzillé G, Chazard E, His-Mahier C, Riou C, Cuggia M

Combining information from a clinical data warehouse and a pharmaceutical database to generate a framework to detect comorbidities in electronic health records

BMC Med Inform Decis Mak 2018 Jan 24;18(1):9

Comorbidities are an increasing healthcare challenge and are important for accurate clinical research. Electronic health records, and clinical data warehouses derived from them, tend to be incomplete with respect to comorbidities. This may be because healthcare organisations are more likely to capture coded diagnoses that are relevant to the services they are providing. However, clinicians are normally keen to ensure they have a complete and up-to-date medication list. This best paper research assessed whether it is possible to use medication lists to identify missing comorbidities, to make a clinical data warehouse more complete and accurate for research use. Sylvestre et al., from the Rennes University Hospital, developed an algorithm that analyses medication lists to identify clinical indications that were not already documented as co-morbidities. For accuracy, the authors only included drug prescriptions with precise indications, by combining data from the French Comorbidity List with the French Theriaque drug database. They additionally selected laboratory tests with very precise indications to identify health conditions that the requesting clinicians must have known about, but which were not listed as coded comorbidities. Their analysis included 4,312 hospital stays with a qualifying prescription. Among the 4,312 patients of the general dataset, 68.4% had at least one drug prescription without a corresponding ICD-10 code. The comorbidity diagnoses suggested by the algorithm were confirmed by experts in 20.3% of reviewed cases. A specialized extract of 122 Ear Nose and Throat hospital stays was used to further evaluate the algorithm, within which comorbidity diagnoses suggested by the algorithm were confirmed in 44.6% of the cases. This research was included as a best paper because it offers and validates a relatively simple approach to semantic enrichment of CDWs, that could be replicated almost anywhere. It highlights the importance of co-morbidity coding, and of poor current practice in such coding.