Clinical Research Informatics: Contributions from 2022

Xavier Tannier[1], **Dipak Kalra**[2], **Section Editors for the IMIA Yearbook Section on Clinical Research Informatics**

[1] Sorbonne University, Inserm, University Sorbonne Paris-Nord, INSERM UMR_S 1142, LIMICS, F-75006 Paris, France

[2] 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 2022.

Method: A bibliographic search using a combination of Medical Subject Headings (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 1,324 papers returned by the search, published in 2022, that were in the scope of the various areas of CRI, the full review process selected four best papers. The first best paper describes the process undertaken in Germany, under the national Medical Informatics Initiative, to define a process and to gain multi-decision-maker acceptance of broad consent for the reuse of health data for research whilst remaining compliant with the European General Data Protection Regulation. The authors of the second-best paper present a federated architecture for the conduct of clinical trial feasibility queries that utilizes HL7 Fast Healthcare Interoperability Resources and an HL7 standard query representation. The third best paper aligns with the overall theme of this Yearbook, the inclusivity of potential participants in clinical trials, with recommendations to ensure greater equity. The fourth proposes a multi-modal modelling approach for large scale phenotyping from electronic health record information. This year's survey paper has also examined equity, along with data bias, and found that the relevant publications in 2022 have focused almost exclusively on the issue of bias in Artificial Intelligence (AI).

Conclusions: The literature relevant to CRI in 2022 has largely been dominated by publications that seek to maximise the reusability of wide scale and representative electronic health record information for research, either as big data for distributed analysis or as a source of information from which to identify suitable patients accurately and equitably for invitation to participate in clinical trials.

Keywords

Observational studies as Topic; real-world data; real-world evidence generation; consent; phenotyping

Yearb Med Inform 2023:146-51
http://dx.doi.org/10.1055/s-0043-1768748

1 Introduction

Within the 2023 International Medical Informatics Association (IMIA) Yearbook, the goal of the Clinical Research Informatics (CRI) section is to provide an overview of research trends from 2022 publications that demonstrate the progress in multi-faceted 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. Although this year's Yearbook focus should ideally embrace human, animal, plant and planetary considerations, we did not find, in 2022, CRI publications that have such a holistic perspective. The selections made have therefore focused on maximising the inclusivity of people within clinical research.

2 About the Paper Selection

A comprehensive review of articles published in 2022 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 twice, initially in November 2022 and then refreshed for late in the year publications in January 2023, considering the electronic publication date. 1,324 papers were found and screened for relevance to CRI, for reporting original research as opposed to opinion or educational papers, and their scientific quality was blindly rated as *low, medium,* or *high* by the two section

editors based on papers' *title* and *abstract*. 55 references classified as high quality contributions to the field by at least one of the two section editors or medium quality by both 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; Feasibility studies, patient recruitment, improved user experiences of CRI systems and Governance (ethical, regulatory, societal, policy issues, stakeholder participation, research networks, team science). A further review by both editors on the basis of full paper screening, again selecting those that scored at least one high or two medium, resulted in nine candidate papers being taken forward to external peer review. In conformance with the IMIA Yearbook process, these nine papers were peer-reviewed by the IMIA Yearbook editors and external reviewers. Each paper was reviewed by at least four reviewers. 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.

3 Outlook

The nine candidate best papers for 2022 illustrate recent efforts and trends in different CRI areas: feasibility studies and patient recruitment; interoperability, data harmonisation and data migration; real-world evidence generation and electronic phenotyping; ethical, legal, and social issues including consent and AI ethics; information security and secure record linkage.

3.1 Feasibility Studies, Patient Recruitment

One of the selected best papers, summarized below, evaluates the reuse of Electronic Health Records (EHRs) for clinical trial feasibility queries across multiple hospital sites [1]. While this capability is already established through exporting hospital data into interoperable data warehouses like OMOP (Observational medical outcomes partnership) or i2b2, the study aims to replicate it using HL7 FHIR. Validation was performed across four German MII hubs. The paper's significance lies in building tools directly on the increasingly adopted FHIR standard, which has the potential to enhance adoption, reduce costs and complexity, and alleviate the need for specialized skills in alternative models.

Table 1 Selection of best papers for the 2023 IMIA Yearbook of Medical Informatics for the Clinical Research Informatics section. The articles are listed in alphabetical order by the first author's surname.

Section
Clinical Research Informatics

- Ahuja Y, Zou Y, Verm, A, Buckeridge D, Li Y. MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record. J Biomed Inform 2022;134:104190. doi: 10.1016/j.jbi.2022.104190.
- Gruendner J, Deppenwiese N, Folz M, Köhler T. Kroll B, Prokosch HU, Rosenau L, Rühle M, Scheidl MA, Schüttler C, Seldmayr B, Twdrik A, Kiel A, Majeed RW. The Architecture of a Feasibility Query Portal for Distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) Patient Data Repositories: Design and Implementation Study. JMIR Med Inform 2022;10(5):e36709. doi: 10.2196/36709.
- Peters U, Turner B, Alvarez D, Murray M, Sharma A, Mohan S, Patel S. Considerations for Embedding Inclusive Research Principles in the Design and Execution of Clinical Trials. Ther Innov Regul Sci 2023;57:186–95. doi: 10.1007/s43441-022-00464-3.
- Zenker S, Strech D, Ihrig, K, Jahns R, Müller G, Schickhardt C, Schmidt G, Speer R, Winkler E, Graf von Kielmansegg S, Drepper J.. Data protection-compliant broad consent for secondary use of health care data and human biosamples for (bio)medical research: Towards a new German national standard. J Biomed Inform 2022;131:104096. doi: 10.1016/j.jbi.2022.104096.

3.2 Interoperability, Data Integration and Harmonization

The challenges of database migration and mapping were examined in a publication on re-designing and migrating an existing rare disease registry for it to be used in regulatory context [2]. This paper focuses on the remodeling of the Registry of Multiple Osteochondromas (REM) to align with regulatory requirements and recommendations. The study emphasizes the key stages involved in achieving semantic interoperability, data quality, and governance within the registry. To enhance interoperability, ontologies and standards were integrated for proper data collection following FAIR principles. Data quality was improved through the addition of parameters and domains, minimizing human error. Furthermore, a two-level governance structure was established to increase visibility for the scientific community and patients. The remodeled REM registry effectively meets the needs of the scientific community and provides valuable real-world evidence for regulatory purposes.

Another strong paper last year addresses the challenge of integrating data across diverse research environments in collaborative projects [3]. While allowing for variation in data collection protocols, achieving interoperability is crucial for successful collaboration. The Clinical Sequence Evidence-Generating Research (CSER) consortium Data Coordinating Center (DCC) shares its experiences in coordinating survey and genomic sequencing data from seven research sites. Fourteen lessons learned and eleven recommendations for future consortia are identified, covering communication and planning, harmonization, informatics, compliance, and analytics. The paper highlights the importance of budgeting for data coordination early in the consortium's development and emphasizes the need for continuous communication and solid data governance approaches.

3.3 Real-world Evidence Generation, Electronic Phenotyping

The routine use of structured electronic healthcare records (EHRs) offers opportunities to address gaps in clinical evi-

dence. However, many publications have underscored the importance of addressing challenges related to verification, validation, data privacy, and the social responsibility of conducting research. To tackle these challenges, the European Society of Cardiology and the BigData@Heart consortium have collaborated with various stakeholders, including patients, clinicians, scientists, regulators, journal editors, and industry experts. They propose the CODE-EHR Minimum Standards Framework [4], which provides a practical approach to enhance study design, transparency, and the effective utilization of healthcare data for research purposes, providing a roadmap for future improvements. This paper contributes to the development of guidelines and best practices that promote the responsible and impactful use of big data in healthcare research.

In the same theme, MixEHR-Guided is described in one of the four selected best papers [5]. It is an automatic phenotyping model that utilizes a probabilistic joint topic modeling approach to projects a patient's high-dimensional and heterogeneous clinical record onto a low-dimensional latent topic mixture membership over disease phenotypes. The training uses a generative latent topic model inspired by Latent Dirichlet Allocation (LDA), but unlike LDA, the model associates topics with reference phenotypes. This allows the result to be interpretable. The authors apply MixEHR-Guided on MIMIC-III and on PopHR (1.3 M patients from Québec, Canada). 1,500 identifiable phenotypic topics are inferred and exhibit meaningful connections among ICD and ATC codes. The inferred phenotypic topics accurately recovered 9 out of 12 rule-based labels. Based on a qualitative comparison with reference phenotypes, the authors claim that MixEHR-G's phenotype topics are clinically meaningful. MixEHR-G is novel in obtaining simultaneous phenotypic predictions (for ~1,500 phenotypes) and improving the interpretability of phenotypic subjects.

3.4 Ethical, Legal, and Social Issues

The highest scoring best paper in the CRI Section presents a model for obtaining broad consent for secondary use of health care data and human bio-samples that is compliant with Europe's data protection Regulation (the GDPR) and has the endorsement of all key ethical and data protection decision makers in Germany [6]. It has up to now been very difficult to gain approval for consent specifications that are relatively broad because the data subject is not deemed to have been well enough informed of the purpose or purposes and organisations processing the data. It is hoped that this example from Germany will inspire other countries to establish a legally acceptable basis for broad consent, which is a critical enabler of big data research ecosystems. Another contrasting ethical paper, also the best paper this year, focuses on the inclusion of diverse populations within clinical trials [7]. This paper explains the importance of study populations genuinely representing the patient populations who will utilise the research results (such as an innovative treatment or an algorithm) and highlight some of the main areas of disparity that are found today amongst clinical trial populations. This paper also positively explains how a proactive approach to equity and representativeness can be taken. The third ethical paper that we highlight deals with compensating statistically for biased will inevitably exist within real-world data [8]. This focuses on the reuse of existing EHR data within a real-world control arm of the study, which is increasingly attractive in rare disease and paediatric research. To be an accurate control the population included from the real-world data must be comparable, not only from the perspective of representativeness of the individuals but best stages in disease and treatment pathways. The authors describe a statistical approach to estimate the bias of the new external control arm, based on previous observed biases. Since real-world data is increasingly available and clinical trials are increasingly expensive, real-world control arms are gaining favour but must be designed robustly. Finally in this ethics sub-section, the literature survey paper is a scoping review on health equity. Maurud *et al.* [9] conducted a scoping review on ways health equity has been promoted in clinical research informatics with patient implications. They found that the literature in 2022 predominantly focused on bias and equity in the data used for AI development. They interestingly found that although there is a risk of algorithm bias through training with biased data, they also note that AI has at times uncovered bias in existing treatment decision-making. AI can therefore be helpful to reduce bias as well as being vulnerable to bias.

3.5 Information Security

Record linkage-based patient intersection cardinality for rare disease studies using Mainzelliste and secure multi-party computation [10]. This paper presents a complementary solution to our best paper on broad consent, to address the challenge of linking patient records from multiple data sources accurately, with the least amount of identifying or potentially identifying data to match the records. The broad consent approach of Zenker *et al.* [6] mentioned above and elaborated below is an approach that can be adopted for pseudonymised data which has a linkage key per patient record. Secure multi-party computation takes the contrasting approach of avoiding the need to incorporate linkage keys within data set, so that each dataset can be processed in an anonymized form, therefore not necessarily requiring consent for this processing. This approach also avoids the need to establish a trusted third-party to perform linkage and then to generate an anonymized data extract for the research users. Most Privacy Preserving Record Linkage solutions utilise a Trusted Third Party (TTP) approach using hashed identifiers to minimise the risk of inappropriate relinkage, but centralised within the TTP the mechanisms that safeguard this process. Secure multi-party computation enables the analysis across multiple data repositories without centralising the data or requiring the centralisation of (hashed) linkage keys. In this paper, the authors extend the Mainzelliste record linkage framework with an error-tolerant, "fuzzy" match algorithm, increasing the linkage quality on noisy and incomplete data.

Acknowledgement

We would like to acknowledge the support of Adrien Ugon, Martina Hutter, Kate Fultz Hollis, Lina Soualmia, 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.

References

1. Gruendner J, Deppenwiese N, Folz M, Köhler T. Kroll B, Prokosch HU, et al. The Architecture of a Feasibility Query Portal for Distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) Patient Data Repositories: Design and Implementation Study. JMIR Med Inform 2022 May 25;10(5):e36709. doi: 10.2196/36709.
2. Mordenti M, Boarini M, D'Alessandro F, Pedrini E, Locatelli M, Sangiorgi L. Remodeling an existing rare disease registry to be used in regulatory context: Lessons learned and recommendations. Front Pharmacol 2022 Sep 23;13:966081. doi: 10.3389/fphar.2022.966081.
3. Muenzen KD, Amendola LM, Kauffman TL, Mittendorf KF, Bensen JT, Chen F, et al. Lessons learned and recommendations for data coordination in collaborative research: The CSER consortium experience. HGG Adv 2022 May 20;3(3):100120. doi: 10.1016/j.xhgg.2022.100120.
4. Kotecha D, Asselbergs FW, Achenbach S, Anker SD, Atar D, Baigent C, et al. CODE-EHR best practice framework for the use of structured electronic healthcare records in clinical research. BMJ 2022 Aug 29;378:e069048. doi: 10.1136/bmj-2021-069048.
5. Ahuja Y, Zou Y, Verm, A, Buckeridge D, Li Y. MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record. J Biomed Inform 2022 Oct;134:104190. doi: 10.1016/j.jbi.2022.104190.
6. Zenker S, Strech D, Ihrig, K, Jahns R, Müller G, Schickhardt C, et al. Data protection-compliant broad consent for secondary use of health care data and human biosamples for (bio)medical research: Towards a new German national standard. J Biomed Inform 2022 Jul;131:104096. doi: 10.1016/j.jbi.2022.104096.
7. Peters U, Turner B, Alvarez D, Murray M, Sharma A, Mohan S, et al. Considerations for Embedding Inclusive Research Principles in the Design and Execution of Clinical Trials. Ther Innov Regul Sci 2023 Mar;57(2):186-95. doi: 10.1007/s43441-022-00464-3.
8. Incerti D, Bretscher MT, Lin R, Harbron C. A meta-analytic framework to adjust for bias in external control studies. Pharm Stat 2023 Jan;22(1):162-80. doi: 10.1002/pst.2266.
9. Maurud S, Henni SH, Moen A. Health Equity in Clinical Research Informatics. Yearb Med Inform 2023 Jul 6:138-45. doi: 10.1055/s-0043-1768720.
10. Kussel T, Brenner T, Tremper G, Schepers J, Lablans M, Hamacher K. Record linkage based patient intersection cardinality for rare disease studies using Mainzelliste and secure multi-party computation. J Transl Med 2022 Oct 8;20(1):458. doi: 10.1186/s12967-022-03671-6.

Correspondence to:
Xavier Tannier
Sorbonne University, Inserm
University Sorbonne Paris-Nord
INSERM UMR_S 1142, LIMICS
F-75006 Paris, France
E-mail: xavier.tannier@sorbonne-universite.fr

Appendix: Summary of Best Papers Selected for the IMIA Yearbook 2023, CRI Section

Ahuja Y, Zou Y, Verm, A, Buckeridge D, Li Y

MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record

J Biomed Inform 2022 Oct;134:104190. doi: 10.1016/j.jbi.2022.104190

It can be challenging to identify a disease cohort within an EHR repository, especially when precision is required to accurately identify patients with a particular disease variant, biomarker specificity or other precisely defined criteria. EHRs may contain a number of working diagnoses that have been assumed during a diagnostic pathway and may contain a number of working diagnoses that have been assumed during a diagnostic pathway and through disease evolution. Electronic phenotyping is a process of defining a set of EHR data item values that, when found together, are highly suggestive of a particular disease, or conversely may establish its absence. This yearbook chapter has included electronic phenotyping algorithms in previous years, but this methodology by Ahuja *et al.* has been included this year because it offers an advance on previous methods. To maximise accuracy, the authors utilise highly dimensional EHR data, and have mapped these to a portfolio of reference phenotypes in order to enhance the efficiency over the classical Latent Dirichlet Allocation (LDA) approach. The authors have modelled around 1,500 phenotype topic maps, and have demonstrated a high performance of matching 1.3 million patients in Québec, Canada to these phenotype topic maps, which can be performed simultaneously. This methodology may therefore advance the ability to construct virtual cohorts and perform precisely targeted big data analyses on heterogeneous health data.

Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch HU, Rosenau L, Rühle M, Scheidl MA, Schüttler C, Sedlmayr B, Twrdik A, Kiel A, Majeed RW

The Architecture of a Feasibility Query Portal for Distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) Patient Data Repositories: Design and Implementation Study

JMIR Med Inform 2022 May 25;10(5):e36709. doi: 10.2196/36709

In this paper, Gruendner and colleagues report the design, implementation and evaluation of a platform with a user query design tool to author clinical trial eligibility criteria as electronic health record queries, and to execute those across a network of hospital electronic health record systems. The reuse of EHRs for clinical trial feasibility is not new, but methods act now have required the export of EHR data into a separate clinical data warehouse, often utilising an OMOP or i2b2 architecture. The authors here have utilised HL7 FHIR and a FHIR-specific standard query formalism. The research incorporates the logical sequence of a review of the literature regarding existing tools and methods, a requirements analysis, an architecture design and implementation, and validation using synthetic data distributed across a number of sites in Germany that are part of the Medical Informatics Initiative. Although utilising a standard query representation, the author and environment offers a non-technical friendly interface for constructing the eligibility queries, guided by an ontology. The advantage of this approach is that it can be executed on FHIR servers, which are growing in popularity as a repository architecture for electronic health record information. It is therefore possible through this methodology for a healthcare organisation such as the hospital to leverage the technologies it already has in place and, quite importantly, the skills and expertise of staff it is more likely to already have in the house, to create an EHR endpoint for these distributed queries.

Peters U, Turner B, Alvarez D, Murray M, Sharma A, Mohan S, Patel S

Considerations for Embedding Inclusive Research Principles in the Design and Execution of Clinical Trials

Ther Innov Regul Sci 2023 Mar;57(2):186-95. doi: 10.1007/s43441-022-00464-3

This paper is, rather unusually, being included is the best paper although it is a review paper. The authors present a very logical and well researched analysis of the disparities and biases in the populations that are recruited into clinical trials, leading to a lack of representativeness of the population from a number of demographic perspectives. They highlight in particular race, ethnicity, socio-economic factors, underserved communities and disparities in the approach to reimbursing trial participants for out-of-pocket expenses directly incurred through the process of trial participation. The authors present evidence of these categories of disparity, and argue for the risk that clinical trial findings will therefore not be broadly applicable to the populations intended to be treated. The paper also goes into mitigations: approaches that can be taken to broader equity and inclusion, to improve the representation within trials of true population diversity. This paper has been included in the yearbook because it presents a clear and convincing case and call to action for all of those involved in clinical research to take measures proactively to ensure clinical trial accessibility and inclusion is equitable and that study populations are genuinely representative.

Zenker S, Strech D, Ihrig K, Jahns R, Müller G, Schickhardt C, Schmidt G, Speer R, Winkler E, von Kielmansegg SG, Drepper J

Data protection-compliant broad consent for secondary use of health care data and human biosamples for (bio)medical research: Towards a new German national standard

J Biomed Inform 2022 Jul;131:104096. doi: 10.1016/j.jbi.2022.104096.

This paper addresses a challenge that every European country, and the EU as a whole, struggles with when seeking to find an acceptable approach to the reuse of health data for research. It is often difficult to robustly anonymise health data. Linkage may be required to join records between multiple

care providers such as a hospital and a GP, and longitudinally to update health record extracts over time. The presence of linkage identifiers means the data is pseudonymised, and normally regarded as personal data under the European General Data Protection Regulation. Some health data types, some uses such as AI development, and the case of rare diseases with small patient numbers or make it difficult to be rigorous in anonymisation. In such cases informed consent is the usual legal basis for processing health data for research. The difficulty is that the GDPR normally requires that consent is fully informed and specific. Health data ecosystems, in contrast, accumulate health data at scale in order to serve multiple future research purposes that cannot be predicted at the time of obtaining consent. The holy grail solution is to seek broad consent from patients to categories of data use, such as categories of research, but so far it has proved challenging for broad consent to be accepted by data protection legislators. The German Medical Informatics programme has undertaken an extensive process of multi-stakeholder consultation, including patient representatives, on how broad consent might be worded and governed, such that it could be acceptable to them and to decision-makers. The consultation process included all 52 German ethics committees and all 18 German Federal and state data protection authorities. It has now been recognised authoritatively as an acceptable process for obtaining broad consent for research using health data. This paper, which was the top ranked in the peer review process, reports on the reasons for pursuing this and the methodology that was adopted in order to obtain the essential endorsements. There are links in the paper to the actual broad consent wording, in English and German. This initiative is the first across Europe to have created a legal and acceptable basis for broad consent and offers a pathway that other countries could now pursue.