Article published online: 2025-04-08

Clinical Research Informatics: a Decade-in-Review

Christel Daniel[1, 2], Peter J. Embí[3]

1 AP-HP, France
2 Sorbonne Université, INSERM UMR_S 1142, LIMICS, F-75006, Paris, France
3 Vanderbilt University Medical Center, Department of Biomedical Informatics, Nashville, Tennessee, USA

Summary

Background: Clinical Research Informatics (CRI) is a subspeciality of biomedical informatics that has substantially matured during the last decade. Advances in CRI have transformed the way clinical research is conducted. In recent years, there has been growing interest in CRI, as reflected by a vast and expanding scientific literature focused on the topic. The main objectives of this review are: 1) to provide an overview of the evolving definition and scope of this biomedical informatics subspecialty over the past 10 years; 2) to highlight major contributions to the field during the past decade; and 3) to provide insights about more recent CRI research trends and perspectives.
Methods: We adopted a modified thematic review approach focused on understanding the evolution and current status of the CRI field based on literature sources identified through two complementary review processes (AMIA CRI year-in-review/IMIA Yearbook of Medical Informatics) conducted annually during the last decade.
Results: More than 1,500 potentially relevant publications were considered, and 205 sources were included in the final review. The review identified key publications defining the scope of CRI and/or capturing its evolution over time as illustrated by impactful tools and methods in different categories of CRI focus. The review also revealed current topics of interest in CRI and prevailing research trends.
Conclusion: This scoping review provides an overview of a decade of research in CRI, highlighting major changes in the core CRI discoveries as well as increasingly impactful methods and tools that have bridged the principles-to-practice gap. Practical CRI solutions as well as examples of CRI-enabled large-scale, multi-organizational and/or multi-national research projects demonstrate the maturity of the field. Despite the progress demonstrated, some topics remain challenging, highlighting the need for ongoing CRI development and research, including the need of more rigorous evaluations of CRI solutions and further formalization and maturation of CRI services and capabilities across the research enterprise.

Keywords

Clinical Research Informatics; Biomedical Research; Clinical Trials; Informatics; Literature Review.

Yearb Med Inform 2024: http://dx.doi.org/10.1055/s-0044-1800732

1. Introduction

Clinical research, whether interventional (*e.g.*, clinical trials) or observational (*e.g.*, quasi-experimental studies, outcomes research), encompasses research conducted on humans or on materials of human origin (*e.g.*, real-world data), and is critical to advancing medical science and public health. Conducting such research is a complex, information-intensive effort, involving multiple actors, workflows, processes, and resources. In this context emerged over 15 years ago the field of Clinical Research Informatics (CRI), a biomedical informatics sub-discipline focused on clinical research that was first defined in 2009 [1]. More recently, several literature reviews have been conducted to further frame and refine the topics of CRI in general [2,3]

The focus of this review is to provide an updated practical and informative overview of the CRI field based on examples of impactful papers selected during the last decade through two regular review processes: the annual "CRI year-in-review" presented by at the Informatics Summit of the American Medical Informatics Association (AMIA year-in-review), and the periodic CRI reviews in the IMIA Yearbook of Medical Informatics (IMIA Yearbook).

Our objectives are threefold:

- to provide an overview of the evolving scope and categories of focus for this subspecialty of biomedical informatics over the past 10 years;
- to highlight major contributions to the CRI field during the last decade in each category and identify efforts by the CRI community to address practical informatics needs for advancing clinical research;
- to provide insights on more recent (*i.e.*, past two years) contributions in CRI and identify current research trends and perspectives for the CRI field.

In the sections that follow, we first describe source selection and information extraction. We then highlight major contributions according to the three dimensions presented above (CRI definition, major CRI contributions, and highlight recent research trends and perspectives).

2. Methods

The authors adopted a modified thematic review approach based, in a pragmatic way, on the analysis of the literature sources resulting from two review processes (AMIA CRI year-in-review/IMIA Yearbook) conducted annually since 2013.

2.1. Research question and scoping questions

The core research question of this review focuses on practically relevant aspects of CRI supporting various steps of the life cycle of clinical research. The authors agreed to consider the following scoping questions: "How has the definition and scope of CRI evolved along time?"; "What are examples of CRI tools and methods that have reached sufficient maturity to support clinical research?"; and "What are the current trends and major challenges for CRI research today?"

2.2. Information sources, search strategy and paper selection

Table 1 summarizes the search strategies used by: (a) Embi in preparation of the "CRI year in review" presented annually at the AMIA Informatics Summit, and (b) by the editors of the CRI section of the IMIA Yearbook. Figure 1 summarizes how the authors identified and evaluated the literature sources shortlisted by both AMIA year-in-review and IMIA Yearbook processes, and then applied those approaches to this review.

2.3. Final paper selection and analysis

The search strategy and paper selection described in the previous section resulted in a total of 345 (164 from the AMIA year-in-review, and 181 from the IMIA Yearbook) scientific publications, and of 327 after removal of duplicates (n=18). After additional literature sources according to the author's expertise and final selection, 205 documents remained on the shortlist included as references for this review paper. Consensus between the two reviewers, who are also the co-authors, was reached by discussion. As with the AMIA and IMIA processes on which this approach was based, the ultimate selection of impactful papers reflects the views of the authors, and is not meant an exhaustive or systematic review *per se*.

3. Results

3.1. CRI definition and categories

3.1.1. CRI definition and scope

In 2009, Embi and Payne proposed a definition of clinical research informatics (CRI) as "the subdomain of biomedical informatics concerned with the development, application, and evaluation of theories, methods, and systems to optimize the design and conduct of clinical research and the analysis, interpretation, and dissemination of the information generated" [1,4] Defined as the intersection of the field of clinical research and the field of biomedical informatics, CRI focuses on the development and evaluation

Table 1. Search strategy and paper selection in both AMIA CRI year-in-review and IMIA Yearbook processes.

1 – Search strategy	
MEDLINE query via PubMed	
(a) AMIA CRI year-in-review	(b) IMIA Yearbook CRI section
(a) Initial search by MeSH terms and keywords, including: "Biomedical Research"[MeSH] AND "Informatics"[MeSH] NOT ("computational biology"[MeSH] OR "genetics"[MeSH]) (b) Limit to (e.g., English, with Abstracts, etc.) Additional articles identified using date restrictions and keyword searches with terms such as: Clinical Trials, Clinical Research, Informatics, Translational, Data Warehouse, Research Registries, Recruitment. Recommendations from colleagues.	Two conceptual axes: clinical research and informatics
	((Biomedical Research[MAJR] OR „Biomedical Research"[TIAB] OR „Clinical research"[TIAB] OR „Medical research"[TIAB] OR Nursing Research[MAJR] OR „Nursing Research"[TIAB] OR „Pharmacovigilance"[TIAB] OR „Patient Selection"[TIAB] OR „phenotyping"[TIAB] OR „genotype-phenotype associations"[TIAB] OR Epidemiologic Research Design[MAJR] OR „Epidemiological Monitoring"[TIAB] OR Evaluation Studies as Topic[MAJR] OR Clinical Studies as Topic[MAJR] OR Multicenter Studies as Topic[MAJR] OR Big data [MAJR] OR „Feasibility Studies"[TIAB] OR "clinical research informatics" [TIAB] OR "eligibility criteria" [TIAB] OR "feasibility criteria" [TIAB] OR "cohort selection" [TIAB] OR "patient recruitment" [TIAB] OR "clinical trial eligibility screening" [TIAB] OR "eligibility determination" [TIAB] OR "patient-trial matching" [TIAB] OR "protocol feasibility" [TIAB] OR "real world evidence" [TIAB] OR data science[MAJR] OR „data science"[TIAB] OR Clinical trial protocols as topic[MAJR])
	AND (medical informatics[MAJR] OR „medical informatics"[TIAB] OR „clinical informatics"[TIAB] OR „medical computer science"[TIAB] OR „medical information science"[TIAB] OR Informatics[MAJR] OR „Informatics"[TIAB] OR „electronic healthcare record"[all fields] OR „electronic health record"[all fields] OR „electronic medical record"[all fields] OR „electronic patient record"[all fields] OR „personal health record"[all fields] OR "big data"[TIAB] OR "real world data" [TIAB] OR (Data Warehousing[MeSH] OR „Data Warehousing"[TIAB] OR „Data Warehouse"[TIAB] OR big data[MeSH] OR „big data"[TIAB]))
	AND (yyyy[DP]) AND English[lang] AND hasabstract[text]
2 – Paper selection	
From the total identified, Embi selected a sampling of impactful papers to be presented in detail, and others to receive brief mentions.	Section editors apply exclusion criteria to the original set of retrieved references to select papers in the scope of CRI and blindly evaluate their contribution based on titles and abstracts. The resulting highly contributive papers are reviewed jointly by the section editors to select a consensual list of approximately 15 candidate best papers representative of all CRI categories. In conformance with the IMIA Yearbook process, these shortlisted papers are peer-reviewed by four IMIA Yearbook editors (the two section editors, and two editors in chief), and external reviewers in order to finally select four papers as best papers.

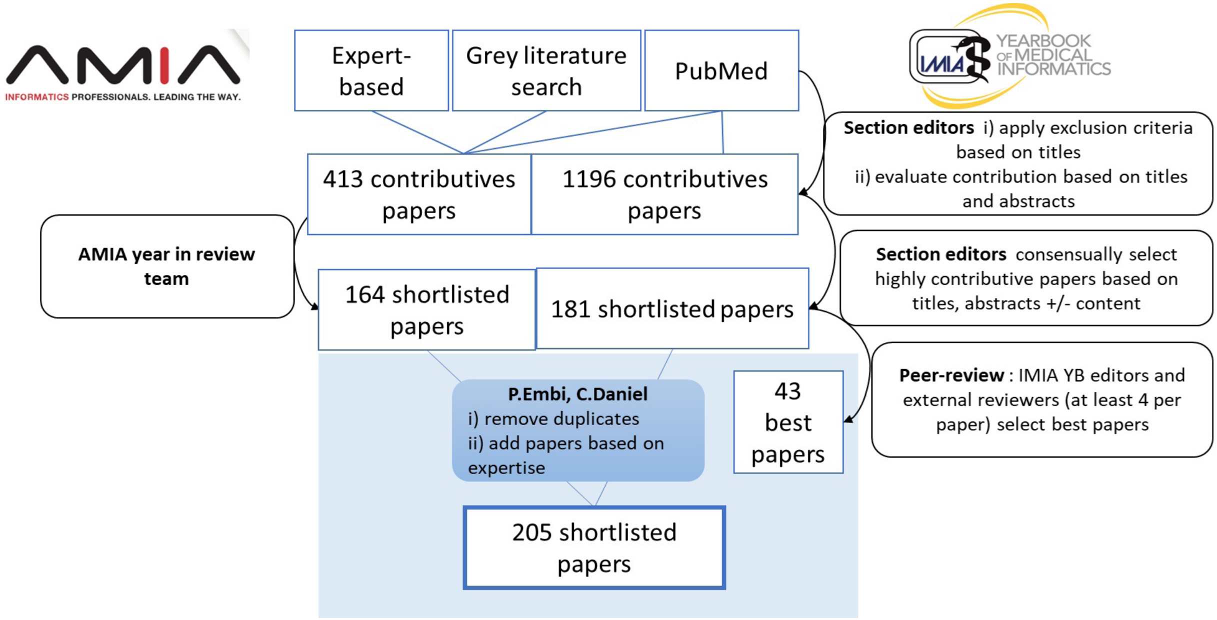
AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.
Expert-based
Grey literature search
PubMed
YEARBOOK OF MEDICAL INFORMATICS
IMIA
Section editors i) apply exclusion criteria based on titles
ii) evaluate contribution based on titles and abstracts
413 contributives papers
1196 contributives papers
AMIA year in review team
Section editors consensually select highly contributive papers based on titles, abstracts +/- content
164 shortlisted papers
181 shortlisted papers
Peer-review : IMIA YB editors and external reviewers (at least 4 per paper) select best papers
P.Embi, C.Daniel
i) remove duplicates
ii) add papers based on expertise
43 best papers
205 shortlisted papers

Figure 1. Overview of the article selection in both AMIA CRI year-in-review and IMIA Yearbook processes, and as applied for purposes of this review (highlighted in the blue box).

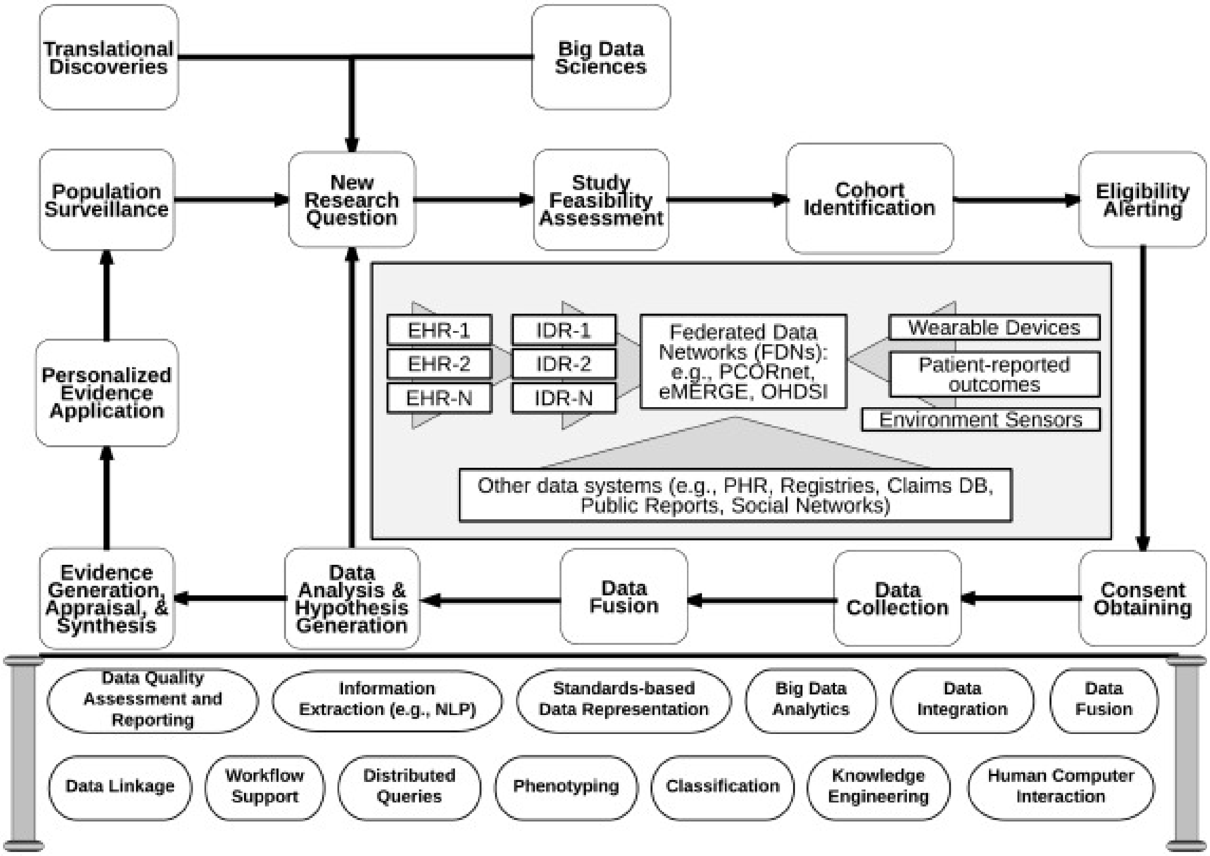
Translational Discoveries
Big Data Sciences
Population Surveillance
New Research Question
Study Feasibility Assessment
Cohort Identification
Eligibility Alerting
Personalized Evidence Application
EHR-1
EHR-2
EHR-N
IDR-1
IDR-2
IDR-N
Federated Data Networks (FDNs): e.g., PCORnet, eMERGE, OHDSI
Wearable Devices
Patient-reported outcomes
Environment Sensors
Other data systems (e.g., PHR, Registries, Claims DB, Public Reports, Social Networks)
Evidence Generation, Appraisal, & Synthesis
Data Analysis & Hypothesis Generation
Data Fusion
Data Collection
Consent Obtaining
Data Quality Assessment and Reporting
Information Extraction (e.g., NLP)
Standards-based Data Representation
Big Data Analytics
Data Integration
Data Fusion
Data Linkage
Workflow Support
Distributed Queries
Phenotyping
Classification
Knowledge Engineering
Human Computer Interaction

Figure 2. Clinical Research Informatics (CRI) conceptual framework [2].

Table 2. Categorization and sub-categorization of the 205 selected papers.

Categories (Acronym, label, number of papers)			Sub categories	Corresponding stage of the clinical research study life cycle	Sources
SD-PR-EXE	Study design & execution Patient recruitment	n=43	Study Design (new CT designs)	CRI step 1: Analyzing Study Designs, feasibility studies	[6–48]
			Patient Recruitment	CRI step 2: Getting Participants into Studies	
			Study Execution (CRI methods & systems, collaborative workflow systems, usability and needs)	CRI step 3: Executing Studies	
PTF	Large scale research platforms	n=12	Study Execution (large scale collaborative platforms, data networks)		[49–60]
DM	Study data management	n=59	Managing Study Data (focus real-world data/phenotyping) (n=17)	CRI step 4: Managing Study Data	[61–77]
			Managing Study Data (focus semantic interoperability, data integration and standardization) (n=19)		[78–96]
			Managing Study Data (focus data quality assessment) (n=14)		[97–110]
			Managing Study Data (focus security, confidentiality) (n=9)		[111–119]
MKD	Data/text mining	n=18	Using Study Data (data or text mining, data visualization)	CRI step 5: Using Study Data	[120–137]
COM-PUB	Results communication and publication	n=5	Communicating Study Results (Research results dissemination)	CRI step 6: Communicating Study Results	[138–141]
			Analyzing Study Publications (knowledge representation, management, or engineering)	CRI step 7: Analyzing Study Publications	[142]
INFRA	Technical infrastructure	n=7	Architecture (Big Data, Cloud computing)		[143–149]
ETH-REG	Ethical & legal issues	n=11	Ethical issues		[150–160]
			Legal, regulatory issues		
SOC	Societal issues	n=27	Societal issues (focus governance, stakeholder participation, engagement)		[161–187]
			Societal issues (focus patient participation, health literacy, patient consenting)		
			Societal issues (focus business model)		
			Societal issues (focus workforce, education and training)		
LHS-SURV	Learning Health Systems Surveillance	n=10	Learning Health Systems, disease, drug & device surveillance		[188–197]
TRE	Trends	n=13	Trends		[1–5,198–205]

of tools and methods to support researchers in clinical research activities including study design, patient recruitment, data collection, integration and analysis. This is further illustrated in the conceptual framework of CRI proposed by Weng and Kahn (Figure 2) [2]

Subsequently, Johnson proposed a search strategy using two conceptual axes defining CRI and sharpening the boundaries with closely related fields such as computational statistics and patient care informatics [3]

As noted by these authors and others, CRI approaches and capabilities have the potential to play a prominent role in supporting widespread advances in medicine, healthcare, and public health [5]

3.1.2. Methodological pillars and categories

Progress by the community of CRI researchers and developers can be depicted as advancing along major pillars of the informatics sub-discipline such as data integration, information extraction, data linkage, data quality assessment, phenotyping, data and text mining, knowledge engineering, and health technology development and assessment (workflow support, human computer interaction, evaluation) in order to reinforce methods and tools supporting the different stages. The papers selected by the AMIA year-in-review and IMIA Yearbook processes have been classified into other sub-categories related to topics and/or steps of the clinical research lifecycle for which they provide solutions [3] . Table 2 summarizes the classification of the final list of selected papers identified via the combined review processes utilized for the current review.

When the literature is plotted by categories over time (as in Figure 3), a different visualization of the evolution of the CRI field is highlighted, with different categories and topics varying by year.

3.2. Main CRI contributions supporting the life cycle of clinical research (2013-2022)

Based on our combined analysis of CRI articles selected over time by the approaches described above, this section provides an overview from 10 years of publications demonstrating progress in the multifaceted aspects of CRI supporting research and innovation in healthcare and biomedicine. Each of the sections that follow cover major categories with exemplars highlighted within.

3.2.1. Study design, participant recruitment and study execution

To support the increasing complexities of clinical research over time, a number of initiatives, including the Human Studies Database [15], the database for aggregate analysis of ClinicalTrials.gov (AACT) [7], the more recent knowledge base of clinical trial eligibility criteria [36], and others [21,35] proposed access to clinical trials repositories allowing significant information to be leveraged and even re-used across studies. In addition, initiatives emerged that were aimed at optimizing clinical research deliverables while meeting the requirements enunciated by the U.S. Food and Drug Administration for electronic systems [8]. Major contributions over the last decade also focused on so-called *secondary use* of health data collected via routine care for optimizing clinical trials [6,9,22,28]. Indeed, many projects and initiatives have explored the value of EHR data for clinical trial design, patient recruitment [12–14,17,19, 25,26,30,31,33,34,37,43,180], or for accelerating data collection [11,39–42]. Facilitated by the generalization of clinical data warehouses, some initiatives were carried out on a large scale within European projects such as the EHR4CR [10,16,20,23,24,29] or TRANSFoRm [27] projects. The EHR4CR project contributed to the development of the InSite commercial solution (formerly Custodix, Belgium, currently part of TriNetX) which reached the evaluation phase [32].

While randomized controlled trials (RCTs) are the gold standard for estimating treatment effects in medical research, there has also been growing interest in and use using real-world data (RWD) for use-cases including drug development, outcomes research, and observational studies. One such use case is the construction of external control arms for evaluation of efficacy in single-arm trials, particularly in cases where randomization is either infeasible or unethical [44,46,47]. In close alignment with a recently released draft framework by the Medical Device Innovation Consortium on real-world clinical evidence and *in vitro* diagnostics, publications report the possibil-

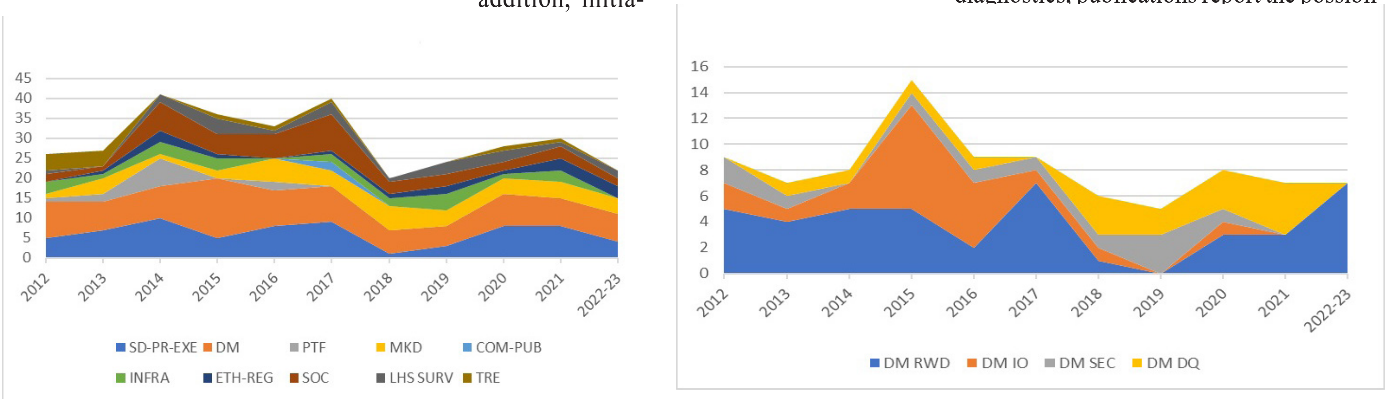

Figure 3. a. Number of papers shortlisted through both AMIA and IMIA review processes (n=327) divided in the ten main categories over time.
b. Number of selected/shortlisted papers in the four sub-categories of data management over time (n=90).

ity of substantially reducing the length and cost of diagnostic test evaluation for diseases with low prevalence to support regulatory decision-making by leveraging subjects from a RWD source [38]. The past decade has also seen the explosion of observational research using EHR data with CRI publications better illuminating our understanding of the specific caveats and considerations/guidelines for conducting observational studies and comparative effectiveness with RWD, while limiting biases [18].

3.2.2. Data management

Real-world data and Clinical Data warehouses

Clinical data warehouses (CDW), built early on at the level of healthcare facilities (Enterprise Data Warehouses) enabled evidence generation based on real-world data to become a reality during the last decade [63,64,67]. In the United States the CDW projects, supported by the Clinical and Translational Science Award (CTSA) funding program of the National Institutes of Health (NIH), have seen increasingly high levels of maturity [62,75]. One successful example is the Informatics for Integrating Biology and the Bedside (i2b2) solution that has been widely used by academic health centers in the U.S. and internationally. Notable i2b2 use-cases include discovery research, case-control studies [66], enable cost-effective genome-wide studies, identify important risks from commonly used medications [61], capture and evaluate patient reported outcomes [48], access image from clinical archive [69], and enable data sharing at scale [70,79,92]. Hospital-based CDWs and multi-site data-sharing activities have also emerged in Europe and Asia over the past decade [65,71,72,76].

Interoperability and standards

Interoperability is another major pilar of CRI that has evolved. In the domain of clinical research, the Clinical Data Information Standards Committee (CDISC) has developed a number of standards for study design, data collection and analysis, and submission to the regulatory bodies [91]. The Operational Data Model (ODM) has been updated to enable the implementation of case report forms supporting more efficiently the complete clinical study lifecycle from the design phase to the collection of patient level data [86]. The complexity and heterogeneity of healthcare data collection and storage approaches has remained a motivating challenge, and there has been movement in the CRI domain toward certain common data standards.

Several research networks focused on cross-institutional data analyzes at scale developed so-called "common data models" (CDM) that require each participating system to transform their underlining data model to the relevant CDM. Examples include the OMOP CDM from the OHDSI consortium [89,94,120], the FDA Mini Sentinel's CDM, I2B2-SHRINE [79], STRIDE or SHARPn [81,143]. New experience has been gained from these projects and the resulting models have been analyzed and compared [88]. While some require transfer of data to a centralized solution, another approach involves federated data and metadata management. One example is the ISO/IEC 11179 standard, that has been demonstrated to semantically link disparate common data elements defined by different organizations [80]. Minimum data sets based on common data elements have been built for different disease areas [56,83,93].

Furthermore, some initiatives have been conducted to bridge standards used in clinical care with clinical research standards, often developed through parallel, and often disparate, efforts. Other harmonization efforts have been conducted within the SHARPn project in mapping Healthcare Clinical Element Models (CEMs) adopted by the Office of the National Coordinator (ONC) to data elements extracted from the CDISC templates [85]. More recently, relying on the increasing adoption of HL7's Fast Healthcare Interoperability Resources (FHIR) by the healthcare industry, many studies have reported the interest of FHIR integration profiles to facilitate the integration of care and research activities [39,92,95]. The HL7 Vulcan acceleration program[1] develops FHIR resources needed to execute prioritized use cases of secondary use of real-world data and especially EHR data. While existing information model standards (*e.g.,* ISO 13606, openEHR, HL7 FHIR, and CDISC) define the basic semantics of health data, semantic interoperability are enabled by medical terminologies, nomenclatures, and ontologies. It is the use of these terminologies through explicitly defined value sets (terminology binding) that brings unambiguous semantics to health data. Consequently, it remains challenging to implement the shift from the "syntactic" level to the "semantic" level of data exchange enabling smart services and applications supporting communication among clinicians, researchers, and healthcare providers. In this context, Semantic Web technologies provide solutions for managing and sharing knowledge and for making the health data FAIR (Findable, Accessible, Interoperable, and Reusable) [87]. In contrast to other major foundational models for clinical research informatics (*e.g.*, BRIDG, CDISC ODM/SDTM, HL7 FHIR, OMOP CDM), the Ontology of Clinical Research (OCRe) takes a logic-oriented ontological modeling approach and attempt to model, not only operational and administrative needs, but also study validity, confounding, and bias needed for assessing study design strength [15]. Beyond the design of ontologies in healthcare (*e.g.*, for epilepsy [78]), research in Semantic Web technologies also includes activities contributing to CRI such as developing efficient ontology editors, and integrating ontologies into healthcare information systems [82,84,90,96]

Data quality assessment

Data quality (DQ) is essential to many CRI activities. Identifying the need for common, standardized approaches to DQ, Kahn *et al.* were motivated to initiate a community-based effort in collaboration with the Electronic Data Methods Forum (EDM) to harmonize terminologies in CRI [99]. This DQ framework has been extended beyond intrinsic aspects to cover

[1] http://hl7.org/fhir/uv/vulcan-rwd/#overview

technical, and contextual categories across the data life cycle enabling the assessment and management of RWD repositories to ensure fitness for purpose. An ontology for characterizing DQ for secondary use of EHR data has been proposed [98]. An increasing number of studies describe DQ assessment methods or tools measuring completeness of data items and other DQ dimensions [100,107,109,110]. Two recent reviews summarize health DQ issues and assessment practices [104,108]. Some initiatives involve patients to assess the quality of their clinical data [103]. Medical centers [102] or large-scale research platforms [101,106] are setting up continuous DQ evaluation and improvement programs.

With the increasing use of observational, non-experimental data for various purposes, the potential for introducing biases and confounding factors that are often hidden in the data must be addressed or at least carefully considered, such as when conducting RWD studies or developing and validating predictive models [74,97,105]. Recent retractions of articles published in high-profile journals reporting COVID-19 studies based on EHR demonstrate that conforming to best practices in developing robust research based on RWD is critical to promote and foster rigor, quality, and reliability of this rapidly growing field [73,140].

Security and data privacy-enhancing techniques

New research opportunities as well as artificial intelligence and machine learning techniques raise the need for greater data access but also bring new potential risks to privacy [113]. The evaluation of this risk in releasing datasets is a major concern [111,112,114]. Many studies focus on de-identification methods and privacy-enhancing techniques [111,115,119], including homomorphic encryption techniques enabling federated analysis of RWD while complying with data protection requirements [117]. An increasing number of papers propose methods for generating synthetic health data [116,118].

Large scale research platforms

Many informatics platforms enabling collaborative research using multi-institutional distributed heterogenous data continued to emerge during the last decade [49]. Large-scale data-sharing initiatives are developing at national level. As examples, in the U.S. the SHARPn project that has proposed a scalable and standards-driven infrastructure for secondary use of EHR data [143]. SHRINE implementations have been used for nationally scalable multi-site disease studies of autism co-morbidity, juvenile idiopathic arthritis, peripartum cardiomyopathy, colorectal cancer, diabetes, and others [79]. The PCORnet initiative links multiple sub-networks via adherence to a common data model. Taken together, PCORnet sites include clinical data from a cohort of 80 million patients and enable rapid access to patient populations for pragmatic clinical trials, epidemiological research, and patient-centered research on both common presentations and rare diseases [53,58].

In Europe, the Health Data Hub in France and the Medical Informatics Initiative in Germany are major national initiatives supporting secondary uses of health data [72,77]. In U.K., we can cite two major initiatives, the Clinical Practice Research Datalink (CPRD) [51] and the ClinicAl disease research using LInked Bespoke studies and Electronic health Records (CALIBER) programs enabling secondary use of nationwide big data from linked electronic health records to improve outcomes in respectively primary care and cardiovascular diseases [57].

At a broad international scale, the Observational Health Data Sciences and Informatics (OHDSI) initiative installed the OMOP common data model as a *de facto* standard for observational research based on large-scale real-word data and demonstrated the value of large-scale data sharing in the analysis of drug use [89]. Finally, during the Covid19 pandemic, national Covid databases were implemented (for example in US (e.g. NC3) [60], Germany or Spain [93]) and international data-sharing platforms were set up (such as 4CE [149] or SCOR [59] projects). In addition, CRI efforts resulted in practical implementations of biobank research systems [50,54].

3.2.3. Data/text mining, artificial intelligence/machine learning, knowledge discovery

One of the most striking findings in CRI over the last decade is the exceptionally rapid development of highly flexible, reusable tools developed to explore, visualize and analyze complex data sets and enable training of artificial intelligence (AI) models that are likely to transform medicine. Growing access to both vast data sets and advanced computational capabilities are enabling the rise of AI technologies and methods for diagnostic or therapeutic decision-support, and other approaches of relevance to CRI. Many studies report the use of machine learning in pharmacovigilance or in precision medicine to predict outcomes such as hospital admissions or in-hospital mortality for specific populations. A significant research effort has emerged to enable federated learning of predictive models from federated data sets [127]. Particular attention is also paid to methodological and organizational issues related to studies using EHR data and to propose methods for bias reduction in studies using EHR data [128,131,133,140].

New infrastructures and methods for data integration/fusion, record linkage, data and text mining have enabled new data discovery opportunities. Significant research efforts in the field focused on the development and validation of phenotyping methods. Rule-based phenotype definitions are collaboratively developed and evaluated within initiatives such as eMERGE [121] or PheKB[2] [122] in U.S. or the CALIBER platform[3] [57], and the phenotype health data gateway[4] in U.K. Many studies exploit the advances in natural language processing (NLP) to extract clinical information from clinical texts [120,126,135]. Others have implemented and promoted innovative approaches for high-throughput phenotyping [81,120,123–125,129,130,132,134].

[2] http://phekb.org

[3] https://www.caliberresearch.org/portal

[4] https://phenotypes.healthdatagateway.org/

Study results communication and publication

Although of undeniable interest in the generation of evidence, especially in areas where traditional clinical trials would be unethical or infeasible, RWD should be reused considering the highest research standards and updated guidelines such as the methodological reference for conducting observational studies developed by European Network of Centers for Pharmacoepidemiology and Pharmacovigilance, ENCePP () under the aegis of the EMA (European Medicines Agency) and the FDA (Food and Drug Administration) or by research communities [140,141]. For example, the CONSORT-AI (Consolidated Standards of Reporting Trials-Artificial Intelligence) extension and its companion statement for clinical trial protocols: SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence) from the EQUATOR network provide recommendations (checklist items) promoting transparency and completeness in reporting clinical trials for AI interventions [205]. These recommendations assist editors and peer reviewers, as well as the general readership, to understand, interpret and critically appraise the quality of clinical trial design and risk of bias in the reported outcomes. In the radiology domain, DECIDE-AI is a new reporting guideline for artificial intelligence studies in radiology. Some other specific concerns of study research communication are e.g., the increase of self-citation [139] or tracking and efficiently displaying research evidences [138,142].

Technical infrastructure

In the current era of massive-scale digitalization of data and computationally-intensive quantitative data analytics, the CRI community has developed new infrastructure capacities that can process, analyze and store petabytes of health data [145,146,148]. The management of massive, unstructured and heterogeneous data is still challenging, especially for supporting the identification of environmental exposures as determinants of physio pathological processes and providing a coherent framework for dealing with multi-scale population data including the phenome, genome, exposome, and their interconnections in the context of the One Health paradigm [65,144,147].

Ethical, legal and societal issues (data governance, patient engagement, sustainability)

A major lesson that the coronavirus disease 2019 (COVID-19) pandemic has taught the scientific CRI community is that healthcare institutions shall continue to strengthen their expertise in data driven evidence generation. Since new opportunities also bring new societal challenges, most of the institutions in charge of CRI and/or CDW for research put in place governance committees of varying forms in charge of setting up the policies of data requests for internal research projects and/or external data sharing agreements as well as reviewing these requests [162]. These governance bodies need to guarantee that the intensive data-driven research is conducted in respect of the law and in line with ethical principles (*e.g.*, that data release fits the purpose of the research), especially in the context of large-scale collaborative research. Ethical and regulatory issues related to the use of large-scale linked data and new AI technologies are especially challenging and need to be addressed to respond to the current rapidly changing data-driven agenda. Researchers have to understand complicated and sometimes contradictory legal requirements inherent to big data projects and to consider ethical obligations in order to balance potential of discovery with legal and ethical considerations [151,152,154,159,160]. Ethics review committees may have to be reformed to deal with big data research and improve their oversight capacity [156,158] and supportive tools (eIRB systems) could be useful in this context. Some publications focus on lessons learned and recommendations related to ethical, legal and societal issues in the specific context of the covid-19 pandemics [157] or rare diseases. In Europe, the European Institute for Innovation through Health Data (i~HD) plays a strong role in many national and international collaborative projects or initiatives involving academic research groups from all over Europe collaborating with global pharmaceutical companies in supporting the ecosystem in especially addressing ethical, legal, societal issues [177]. Key national initiatives are promoting CRI [161,166,171,172,178,179,181]. Significant research efforts in the CRI field also focused on specific societal issues : patient information and consent [150,153,155,168,182], patient involvement in the design and execution of data-intensive medical research and clinical trials [164,170,175,176,183], health equity in CRI [180,184,186], sustainability [68,163], the role of IT vendors [167,185], barriers to data sharing [165,173] and more generally educational issues [169,174].

Learning Health System and large-scale surveillance

CRI contributions have also advanced support for the realization of a learning health system that enables healthcare systems to objectively and continuously monitor their practices, pragmatically compare the outcomes of care over time, including patient-reported outcomes, and simulate and evaluate the impacts of certain strategic decisions on organizations or the quality of care [191,192,194,196]. Major research efforts also focused on data-driven approaches for signal detection in pharmacovigilance especially in post marketing drug and vaccine safety surveillance and more generally patient safety [188–190,193,195].

3.3. Recent (last two years) research trends and perspectives (2022-2023)

3.3.1. Health technology assessment, maturity of CRI services and innovative clinical trial designs

Despite considerable progress over the past 10 plus years, certain challenges remain a focus for CRI now and into the future. CRI interventions still appear too often as anecdotes or quasi-experimental studies, whereas methodologically rigorous studies (*e.g.*, randomized, controlled studies) of CRI interventions remain rare. As the field matures, more efforts will be needed to demonstrate the successes and value of CRI interventions. Moreover, there is a need to enhance the level of formalization of CRI services in large academic centers. ***Ensuring high-quality research*** relies on multiple factors including variations in organizational culture of academic medical centers, retention of research workforces, efficient multi-departmental organizational collaborations (e.g., relationship between biomedical research expertise and enterprise IT), technical skills, operational efficiencies and transparent communication of standard policies, procedures and key performance indicators [199].

Another emerging concern relevant to CRI is the development, validation, and evaluation of ***AI-based healthcare technology***. In the context of medical AI, that survived several "winters" since the 1970's, there is a need to better understand how new AI technologies will challenge current strategies for regulating and validating AI devices for medicine and research. Recent reviews focusing on the development of AI in medicine state that AI could transform clinician workflows, patient care, administrative tasks, and augment medical knowledge and decision making [203,204]. They also underline unsolved challenges, from issues about training ML systems to unclear accountability making AI's implementation difficult and largely ununderstood by healthcare professionals still struggling to implement AI in their daily practice. Guidelines for clinical trial protocols for interventions involving AI are available and some papers propose AI implementation frameworks and approaches for AI development, implementation, use, monitoring and governance [131,133,198].

As the regulatory environment becomes progressively receptive to utilizing real-world evidence in clinical trials, many publications focus on various ***innovative hybrid clinical trial settings***. Of particular interest is the use of RWD to accelerate and perhaps even approximate certain traditional research approaches. Among innovative trial designs, platform trials including multiple trial arms (conducted simultaneously and/or sequentially) on different treatments under a single master protocol. The capabilities of medical centers to perform EHR-based protocol feasibility assessment, clinical site selection, and patient pre-screening are keys for successful platform trials [45]. An increasing number of initiatives, such as the TAES (TriAl Eligibility Surveillance) system, leverage existing real-world data, using artificial intelligence and standard data models, to enhance identification of patients potentially eligible for clinical trials and generate timely notifications to physicians and simultaneously decrease the burden on research teams of manual EHR review [43]. Target trial emulation can be used for eligibility determination [200] and also to assess agreement between clinical trials and analogue real-world evidence studies. Concordance in results vary depending on the agreement metric. Emulation differences, chance, and residual confounding can contribute to divergence in results [201].

3.3.2. Translational research, better integration of care and research and Learning Health Systems

There is a need to promote organizational and technical interoperability between care and research domains towards the implementation of "***clinical trials as care option***". There are an increasing number of initiatives aiming at avoiding double data entry through the integration of electronic case report forms (eCRF) systems with electronic health records (EHR) – the "***eSource" or "EHR2EDC" use case***. While technical capabilities are important, organizational priorities, structure, and the site's support of clinical research functions are equally important considerations. According to a recent study assessing site readiness for eSource, only 21% of sites were using Fast Healthcare Interoperability Resources (FHIR) standards to exchange patient data with other institutions. Respondents generally gave lower readiness for change ratings to organizations that did not have a separate research information technology group and where researchers practiced in hospitals not operated by their medical schools [41]. Although a recent study demonstrated that secure, consistent, and automated end-to-end data transmission from the treating physician to the regulatory authority was feasible, the industry-wide implementation of EHR2EDC still requires policy decisions that set the framework for the use of research data based on routine care data [42]. To overcome the difficulty of integration with EHR systems due to information security and data protection restrictions, a novel architecture for a federated electronic data capture system (fEDC) has been proposed [202].

Despite recent efforts, especially those led by HL7 organization, for example the ***HL7 FHIR Vulcan accelerator***, formal representation of multimodal and multi-level data supporting data interoperability across clinical research and care domains is still challenging. Additional efforts are required from standard development organizations (CDISC, HL7 and DICOM) to better align the standards used in care and research domain and from the vendors of clinical and biomedical research applications to implement interoperability by design in the development of their solutions within a "single source" approach. Common data models solve many challenges of standardizing electronic health records (EHR) data but are unable to semantically integrate all of the resources [96]. Moreover, semantic interoperability enabling secondary use of clinical data requires a rigorous governance process to ensure internationally the quality of the data standardization process consistently across care and research domains. Recognizing that interoperability frameworks must be continuously adapted to the user's needs and that the update can hardly be fully automatized, a clear collaborative process is needed to efficiently support the creation of new semantic resources scoped to any additional use case of biomedical research.

The persistent challenges posed by data collection based on paradigms of ***primary use of data vs. secondary use*** remain a threat to achieving high quality/value data at the source. Currently considerable resources are dedicated to secondary use of health data and to complex pipelines curating and enriching data outside of the scope of patient care. Efforts to promote improved data quality by design in clinical settings will result in higher quality data for research and evidence generation. For ethical reasons and for supporting the implementation of the ***Learning Heath System,*** ensuring the quality as well as the privacy-preserving accessibility of clinical data for primary use should be considered a priority with major impacts on secondary use and the research agenda.

3.3.3. Trends related to the technological pillar

Progress is being made in ***data quality management*** systems standardizing medical data extraction and quality control methods [108–110]. The development of methods and tools to promote the use of ***textual data and advanced NLP methods*** in real-world observational studies is still an active research area with important contributions from such groups as the NLP Working Group at the Observational Health Data Sciences and Informatics (OHDSI) consortium[135]. Major research efforts are also devoted to ***phenotyping***. Clinical experts and computer scientists still experience a variety of challenges when building phenotypes, including challenges to discerning key clinical events, clinical reasoning, and temporal elements [136]. Standard representations for phenotypes are increasingly being promoted to support the creation, localization and sharing of highly specific phenotype algorithms [96,137]. Additional research is also needed for building secure and efficient infrastructures for "Big data" management supporting, for instance, federated machine learning on distributed data sets hosted on sovereign cloud environments [127].

3.3.4. Trends related to the ethical, legal and societal issues (ELSI)

Ethical, legal and societal research continues to be needed to define and ensure an ethically-guided, fair and patient-centric approach for CRI. Such work is essential to better understand and solve practical issues related to potential biases and inequities arising from the use of RWD and AI [186,187]. Specifically, strategies for appropriate constituent engagement are essential to ensure that solutions like AI can meaningfully benefit patients and other end users. As progress continues toward a new era of digital health driven by cloud data storage, distributed computing, and machine learning, key stakeholders must collaborate to ensure research findings and their downstream implementations are inclusive, representative and accessible for all who should benefit, and work in CRI is key to this.

4. Conclusions

As stated above, our goal was to provide an updated practical and informative overview of the CRI field via a scoping reviewing of CRI publications over the last decade. By leveraging two annual review processes, the "AMIA CRI year-in-review" and IMIA yearbook approach, we started with over 1,500 publications, and ultimately selected 205 impactful papers to analyze in-depth for the current review.

The review results provide an overview of (a) the evolving scope of CRI activities based on, (b) a decade of research in CRI, highlighting major evaluations in core CRI discoveries, innovations, and practical CRI methods and tools, and (c) more recent (past 2 years) trends and areas of CRI emphasis. The findings, as well as examples of enabled large scale research projects across institutions, regionally, nationally and internationally, demonstrate a continued maturation of the field. The findings also show that some important topics remain in focus, reflecting ongoing challenges in CRI that have not yet been fully addressed. They also reveal a need for improved and more rigorous evaluation of CRI solutions and socio-technical interventions. More research is also needed to better integrate the care and research workflows and to implement and realize the goals of a Learning Heath Systems. Similarly, increased focus on building and evaluating robust technical, governance, ethical and policy approaches is essential to build and operate the secure, effective, equitable and efficient infrastructures needed for improving clinical and translational research in an increasingly data-rich, interconnected, and computationally powerful environments. As the field continues to mature, the trajectory of progress and impact that CRI has had over the past decade can be expected to continue, and that should accrue benefits for all.

Acknowledgements

Both authors contributed to the conception, drafting and major editing of the manuscript. The authors have no relevant disclosures to report.

References

1. Embi PJ, Payne PRO. Clinical research informatics: challenges, opportunities and definition for an emerging domain. J Am Med Inform Assoc JAMIA 2009;16:316–27. https://doi.org/10.1197/jamia.M3005.
2. Weng C, Kahn MG. Clinical Research Informatics for Big Data and Precision Medicine. Yearb Med Inform 2016:211–8. https://doi.org/10.15265/IY-2016-019.
3. Johnson SB. Clinical Research Informatics: Supporting the Research Study Lifecycle. Yearb Med Inform 2017;26:193–200. https://doi.org/10.15265/IY-2017-022.
4. Embi PJ. Clinical research informatics: survey of recent advances and trends in a maturing field. Yearb Med Inform 2013;8:178–84.
5. Solomonides A. Review of Clinical Research Informatics. Yearb Med Inform 2020;29:193–202. https://doi.rg/10.1055/s-0040-1701988.
6. Murphy SN, Dubey A, Embi PJ, Harris PA, Richter BG, Turisco F, *et al.* Current state of information technologies for the clinical research enterprise across academic medical centers. Clin Transl Sci 2012;5:281–4. https://doi.org/10.1111/j.1752-8062.2011.00387.x.
7. Tasneem A, Aberle L, Ananth H, Chakraborty S, Chiswell K, McCourt BJ, *et al.* The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. PloS One 2012;7:e33677. https://doi.org/10.1371/journal.pone.0033677.
8. Bansal A, Chamberlain R, Karr S, Kwasa S, McLaughlin B, Nguyen B, et al. A 21 CFR Part 11 compliant graphically based electronic system for clinical research documentation. J Med Syst 2012;36:1661–72. https://doi.org/10.1007/s10916-010-9627-4.
9. Marsolo K. Informatics and operations-let's get integrated. J Am Med Inform Assoc JAMIA 2013;20:122–4. https://doi.org/10.1136/amiajnl-2012-001194.
10. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med 2013;274:547–60. https://doi.org/10.1111/joim.12119.
11. Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch H-U, *et al.* Secondary use of routine-

ly collected patient data in a clinical trial: an evaluation of the effects on patient recruitment and data acquisition. Int J Med Inf 2013;82:185–92. https://doi.org/10.1016/j.ijmedinf.2012.11.008.

12. Weng C, Li Y, Berhe S, Boland MR, Gao J, Hruby GW, *et al.* An Integrated Model for Patient Care and Clinical Trials (IMPACT) to support clinical research visit scheduling workflow for future learning health systems. J Biomed Inform 2013;46:642–52. https://doi.org/10.1016/j.jbi.2013.05.001.
13. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc JAMIA 2014;21:221–30. https://doi.org/10.1136/amiajnl-2013-001935.
14. Fernández-Breis JT, Maldonado JA, Marcos M, Legaz-García M del C, Moner D, Torres-Sospedra J, *et al.* Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. J Am Med Inform Assoc JAMIA 2013;20:e288-296. https://doi.org/10.1136/amiajnl-2013-001923.
15. Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, *et al.* The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research. J Biomed Inform 2014;52:78–91. https://doi.org/10.1016/j.jbi.2013.11.002.
16. Doods J, Botteri F, Dugas M, Fritz F, EHR4CR WP7. A European inventory of common electronic health record data elements for clinical trial feasibility. Trials 2014;15:18. https://doi.org/10.1186/1745-6215-15-18.
17. Schreiweis B, Trinczek B, Köpcke F, Leusch T, Majeed RW, Wenk J, *et al.* Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. Int J Med Inf 2014;83:860–8. https://doi.org/10.1016/j.ijmedinf.2014.08.005.
18. Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. Yearb Med Inform 2014;9:215–23. https://doi.org/10.15265/IY-2014-0009.
19 Miotto R, Weng C. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. J Am Med Inform Assoc JAMIA 2015;22:e141-150. https://doi.org/10.1093/jamia/ocu050.
20. Soto-Rey I, Trinczek B, Girardeau Y, Zapletal E, Ammour N, Doods J, *et al.* Efficiency and effectiveness evaluation of an automated multi-country patient count cohort system. BMC Med Res Methodol 2015;15:44. https://doi.org/10.1186/s12874-015-0035-9.
21. Wu DTY, Hanauer DA, Mei Q, Clark PM, An LC, Proulx J, *et al.* Assessing the readability of ClinicalTrials.gov. J Am Med Inform Assoc JAMIA 2016;23:269–75. https://doi.org/10.1093/jamia/ocv062.
22. Dugas M. Clinical Research Informatics: Recent Advances and Future Directions. Yearb Med Inform 2015;10:174–7. https://doi.org/10.15265/IY-2015-010.
23. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, *et al.* Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform 2015;53:162–73. https://doi.org/10.1016/j.jbi.2014.10.006.
24. Beresniak A, Schmidt A, Proeve J, Bolanos E, Patel N, Ammour N, *et al.* Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the Electronic Health Records for Clinical Research (EHR4CR) European Project. Contemp Clin Trials 2016;46:85–91. https://doi.org/10.1016/j.cct.2015.11.011.
25. Ateya MB, Delaney BC, Speedie SM. The value of structured data elements from electronic health records for identifying subjects for primary care clinical trials. BMC Med Inform Decis Mak 2016;16:1. https://doi.org/10.1186/s12911-016-0239-x.
26. Eubank MH, Hyman DM, Kanakamedala AD, Gardos SM, Wills JM, Stetson PD. Automated eligibility screening and monitoring for genotype-driven precision oncology trials. J Am Med Inform Assoc JAMIA 2016;23:777–81. https://doi.org/10.1093/jamia/ocw020.
27. Mastellos N, Bliźniuk G, Czopnik D, McGilchrist M, Misiaszek A, Bródka P, *et al.* Feasibility and acceptability of TRANSFoRm to improve clinical trial recruitment in primary care. Fam Pract 2016;33:186–91. https://doi.org/10.1093/fampra/cmv102.
28. Hruby GW, Hoxha J, Ravichandran PC, Mendonça EA, Hanauer DA, Weng C. A data-driven concept schema for defining clinical research data needs. Int J Med Inf 2016;91:1–9. https://doi.org/10.1016/j.ijmedinf.2016.03.008.
29. Girardeau Y, Doods J, Zapletal E, Chatellier G, Daniel C, Burgun A, *et al.* Leveraging the EHR4CR platform to support patient inclusion in academic studies: challenges and lessons learned. BMC Med Res Methodol 2017;17:36. https://doi.org/10.1186/s12874-017-0299-3.
30. Krischer J, Cronholm PF, Burroughs C, McAlear CA, Borchin R, Easley E, *et al.* Experience With Direct-to-Patient Recruitment for Enrollment Into a Clinical Trial in a Rare Disease: A Web-Based Study. J Med Internet Res 2017;19:e50. https://doi.org/10.2196/jmir.6798.
31. Obeid JS, Beskow LM, Rape M, Gouripeddi R, Black RA, Cimino JJ, *et al.* A survey of practices for the use of electronic health records to support research recruitment. J Clin Transl Sci 2017;1:246–52. https://doi.org/10.1017/cts.2017.301.
32. 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. https://doi.org/10.1016/j.jbi.2018.12.004.
33. 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. https://doi.org/10.1016/j.ijmedinf.2019.05.018.
34. 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 JAMIA 2019;26:1209–17. https://doi.org/10.1093/jamia/ocz168.
35. Kim JH, Ta CN, Liu C, Sung C, Butler AM, Stewart LA, *et al.* Towards clinical data-driven eligibility criteria optimization for interventional COVID-19 clinical trials. J Am Med Inform Assoc JAMIA 2021;28:14–22. https://doi.org/10.1093/jamia/ocaa276.
36. Liu H, Chi Y, Butler A, Sun Y, Weng C. A knowledge base of clinical trial eligibility criteria. J Biomed Inform 2021;117:103771. https://doi.org/10.1016/j.jbi.2021.103771.
37. 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;119:103822. https://doi.org/10.1016/j.jbi.2021.103822.
38. Chen W-C, 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;31:375–90. https://doi.org/10.1080/10543406.2021.1877724.
39. 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;121:103871. https://doi.org/10.1016/j.jbi.2021.103871.
40. Cheng AC, Banasiewicz MK, Johnson JD, Sulieman L, Kennedy N, Delacqua F, *et al.* Evaluating automated electronic case report form data entry from electronic health records. J Clin Transl Sci 2023;7:e29. https://doi.org/10.1017/cts.2022.514.
41. Eisenstein EL, Zozus MN, Garza MY, Lanham HJ, Adagarla B, Walden A, *et al.* Assessing clinical site readiness for electronic health record (EHR)-to-electronic data capture (EDC) automated data collection. Contemp Clin Trials 2023;128:107144. https://doi.org/10.1016/j.cct.2023.107144.
42. Mueller C, Herrmann P, Cichos S, Remes B, Junker E, Hastenteufel T, *et al.* Automated Electronic Health Record to Electronic Data Capture Transfer in Clinical Studies in the German Health Care System: Feasibility Study and Gap Analysis. J Med Internet Res 2023;25:e47958. https://doi.org/10.2196/47958.
43. Meystre SM, Heider PM, Cates A, Bastian G, Pittman T, Gentilin S, *et al.* Piloting an automated clinical trial eligibility surveillance and provider alert system based on artificial intelligence and standard data models. BMC Med Res Methodol 2023;23:88. https://doi.org/10.1186/s12874-023-01916-6.
44. Lin J, Yu G, Gamalo M. Matching within a hybrid RCT/RWD: framework on associated causal estimands. J Biopharm Stat 2023;33:439–51. https://doi.org/10.1080/10543406.2022.2105346.
45. Lombardo G, Couvert C, Kose M, Begum A, Spiertz C, Worrell C, *et al.* Electronic health records (EHRs) in clinical research and platform trials: Application of the innovative EHR-based methods developed by EU-PEARL. J Biomed Inform

2023;148:104553. https://doi.org/10.1016/j.jbi.2023.104553.

46. Incerti D, Bretscher MT, Lin R, Harbron C. A meta-analytic framework to adjust for bias in external control studies. Pharm Stat 2023;22:162–80. https://doi.org/10.1002/pst.2266.
47. 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;107:369–77. https://doi.org/10.1002/cpt.1586.
48. Pfiffner PB, Pinyol I, Natter MD, Mandl KD. C3-PRO: Connecting ResearchKit to the Health System Using i2b2 and FHIR. PloS One 2016;11:e0152722. https://doi.org/10.1371/journal.pone.0152722.
49. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, *et al.* A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care 2012;50 Suppl:S49-59. https://doi.org/10.1097/MLR.0b013e318259c02b.
50. Bledsoe MJ, Grizzle WE, Clark BJ, Zeps N. Practical implementation issues and challenges for biobanks in the return of individual research results. Genet Med Off J Am Coll Med Genet 2012;14:478–83. https://doi.org/10.1038/gim.2011.67.
51. O'Meara H, Carr DF, Evely J, Hobbs M, McCann G, van Staa T, *et al.* Electronic health records for biological sample collection: feasibility study of statin-induced myopathy using the Clinical Practice Research Datalink. Br J Clin Pharmacol 2014;77:831–8. https://doi.org/10.1111/bcp.12269.
52. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, *et al.* PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc JAMIA 2014;21:602–6. https://doi.org/10.1136/amiajnl-2014-002743.
53. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform AssocJAMIA 2014;21:578–82. https://doi.org/10.1136/amiajnl-2014-002747.
54. van Ommen G-JB, Törnwall O, Bréchot C, Dagher G, Galli J, Hveem K, et al. BBMRI-ERIC as a resource for pharmaceutical and life science industries: the development of biobank-based Expert Centres. Eur J Hum Genet EJHG 2015;23:893–900. https://doi.org/10.1038/ejhg.2014.235.
55. Kass-Hout TA, Xu Z, Mohebbi M, Nelsen H, Baker A, Levine J, *et al.* OpenFDA: an innovative platform providing access to a wealth of FDA's publicly available data. J Am Med Inform Assoc JAMIA 2016;23:596–600. https://doi.org/10.1093/jamia/ocv153.
56. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. J Am Med Inform Assoc JAMIA 2018;25:32–9. https://doi.org/10.1093/jamia/ocx084.
57. 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 JAMIA 2019;26:1545–59. https://doi.org/10.1093/jamia/ocz105.
58. 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;129:60–7. https://doi.org/10.1016/j.jclinepi.2020.09.036.
59. Raisaro JL, Marino F, Troncoso-Pastoriza J, Beau-Lejdstrom R, Bellazzi R, Murphy R, *et al.* SCOR: A secure international informatics infrastructure to investigate COVID-19. J Am Med Inform Assoc JAMIA 2020;27:1721–6. https://doi.org/10.1093/jamia/ocaa172.
60. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, *et al.* The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc JAMIA 2021;28:427–43. https://doi.org/10.1093/jamia/ocaa196.
61. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc JAMIA 2012;19:181–5. https://doi.org/10.1136/amiajnl-2011-000492.
62. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc JAMIA 2012;19:e119-124. https://doi.org/10.1136/amiajnl-2011-000508.
63. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13:395–405. https://doi.org/10.1038/nrg3208.
64. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, *et al.* Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013;51:S30-37. https://doi.org/10.1097/MLR.0b013e31829b1dbd.
65. Martin Sanchez F, Gray K, Bellazzi R, Lopez-Campos G. Exposome informatics: considerations for the design of future biomedical research information systems. J Am Med Inform Assoc JAMIA 2014;21:386–90. https://doi.org/10.1136/amiajnl-2013-001772.
66. Johnson EK, Broder-Fingert S, Tanpowpong P, Bickel J, Lightdale JR, Nelson CP. Use of the i2b2 research query tool to conduct a matched case-control clinical research study: advantages, disadvantages and methodological considerations. BMC Med Res Methodol 2014;14:16. https://doi.org/10.1186/1471-2288-14-16.
67. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, *et al.* Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014;52:28–35. https://doi.org/10.1016/j.jbi.2014.02.003.
68. Wilhelm EE, Oster E, Shoulson I. Approaches and costs for sharing clinical research data. JAMA 2014;311:1201–2. https://doi.org/10.1001/jama.2014.850.
69. Murphy SN, Herrick C, Wang Y, Wang TD, Sack D, Andriole KP, *et al.* High throughput tools to access images from clinical archives for research. J Digit Imaging 2015;28:194–204. https://doi.org/10.1007/s10278-014-9733-9.
70. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc JAMIA 2016;23:909–15. https://doi.org/10.1093/jamia/ocv188.
71. Jannot A-S, Zapletal E, Avillach P, Mamzer M-F, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. Int J Med Inf 2017;102:21–8. https://doi.org/10.1016/j.ijmedinf.2017.02.006.
72. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics Initiatives: Two National Projects to Promote Data Sharing in Healthcare. Yearb Med Inform 2019;28:195–202. https://doi.rg/10.1055/s-0039-1677917.
73. Kotecha D, Asstvelbergs FW, Achenbach S, Anker SD, Atar D, Baigent C, *et al.* CODE-EHR best-practice framework for the use of structured electronic health-care records in clinical research. Lancet Digit Health 2022;4:e757–64. https://doi.org/10.1016/S2589-7500(22)00151-0.
74. Ostropolets A, Albogami Y, Conover M, Banda JM, Baumgartner WA, Blacketer C, *et al.* Reproducible variability: assessing investigator discordance across 9 research teams attempting to reproduce the same observational study. J Am Med Inform Assoc JAMIA 2023;30:859–68. https://doi.org/10.1093/jamia/ocad009.
75. Knosp BM, Dorr DA, Campion TR. Maturity in enterprise data warehouses for research operations: Analysis of a pilot study. J Clin Transl Sci 2023;7:e70. https://doi.org/10.1017/cts.2023.23.
76. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, *et al.* Trustworthy reuse of health data: a transnational perspective. Int J Med Inf 2013;82:1–9. https://doi.org/10.1016/j.ijmedinf.2012.11.003.
77. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018;57:e50–6. https://doi.org/10.3414/ME18-03-0003.
78. Sahoo SS, Lhatoo SD, Gupta DK, Cui L, Zhao M, Jayapandian C, *et al.* Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care. J Am Med Inform Assoc JAMIA 2014;21:82–9. https://doi.org/10.1136/amiajnl-2013-001696.
79. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, *et al.* SHRINE: enabling nationally scalable multi-site disease studies. PloS One 2013;8:e55811. https://doi.org/10.1371/journal.pone.0055811.
80. Sinaci AA, Laleci Erturkmen GB. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. J Biomed Inform 2013;46:784–94. https://doi.org/10.1016/j.jbi.2013.05.009.
81. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, *et al.* Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc JAMIA 2013;20:e341-348. https://doi.org/10.1136/

amiajnl-2013-001939.
82. Liang SF, Taweel A, Miles S, Kovalchuk Y, Spiridou A, Barratt B, *et al.* Semi automated transformation to OWL formatted files as an approach to data integration. A feasibility study using environmental, disease register and primary care clinical data. Methods Inf Med 2015;54:32–40. https://doi.org/10.3414/ME13-02-0029.
83. Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. J Am Med Inform Assoc JAMIA 2015;22:76–85. https://doi.org/10.1136/amiajnl-2014-002794.
84. Pang C, Hendriksen D, Dijkstra M, van der Velde KJ, Kuiper J, Hillege HL, et al. BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. J Am Med Inform Assoc JAMIA 2015;22:65–75. https://doi.org/10.1136/amiajnl-2013-002577.
85. Jiang G, Evans J, Oniki TA, Coyle JF, Bain L, Huff SM, et al. Harmonization of detailed clinical models with clinical study data standards. Methods Inf Med 2015;54:65–74. https://doi.org/10.3414/ME13-02-0019.
86. Huser V, Sastry C, Breymaier M, Idriss A, Cimino JJ. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). J Biomed Inform 2015;57:88–99. https://doi.org/10.1016/j.jbi.2015.06.023.
87. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
88. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016;64:333–41. https://doi.org/10.1016/j.jbi.2016.10.016.
89. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A 2016;113:7329–36. https://doi.org/10.1073/pnas.1510502113.
90. Alonso-Calvo R, Paraiso-Medina S, Perez-Rey D, Alonso-Oset E, van Stiphout R, Yu S, et al. A semantic interoperability approach to support integration of gene expression and clinical data in breast cancer. Comput Biol Med 2017;87:179–86. https://doi.org/10.1016/j.compbiomed.2017.06.005.
91. Yamamoto K, Ota K, Akiya I, Shintani A. A pragmatic method for transforming clinical research data from the research electronic data capture “REDCap” to Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM): Development and evaluation of REDCap2SDTM. J Biomed Inform 2017;70:65–76. https://doi.org/10.1016/j.jbi.2017.05.003.
92. 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.
93. 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;115:103697. https://doi.org/10.1016/j.jbi.2021.103697.
94. Phuong J, Zampino E, Dobbins N, Espinoza J, Meeker D, Spratt H, *et al.* Extracting Patient-level Social Determinants of Health into the OMOP Common Data Model. AMIA Annu Symp Proc AMIA Symp 2021;2021:989–98.
95. Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch H-U, *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;10:e36709. https://doi.org/10.2196/36709.
96. Callahan TJ, Stefanski AL, Wyrwa JM, Zeng C, Ostropolets A, Banda JM, *et al.* Ontologizing health systems data at scale: making translational discovery a reality. NPJ Digit Med 2023;6:89. https://doi.org/10.1038/s41746-023-00830-x.
97. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak 2014;14:51. https://doi.org/10.1186/1472-6947-14-51.
98. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Application of An Ontology for Characterizing Data Quality For a Secondary Use of EHR Data. Appl Clin Inform 2016;7:69–88. https://doi.org/10.4338/ACI-2015-08-RA-0107.
99. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, *et al.* A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS Wash DC 2016;4:1244. https://doi.org/10.13063/2327-9214.1244.
100. Estiri H, Stephens KA, Klann JG, Murphy SN. Exploring completeness in clinical data research networks with DQe-c. J Am Med Inform Assoc JAMIA 2018;25:17–24. https://doi.org/10.1093/jamia/ocx109.
101. 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 Wash DC 2018;6:3. https://doi.org/10.5334/egems.199.
102. 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 2019;181:104804. https://doi.org/10.1016/j.cmpb.2018.10.016.
103. Bell SK, Delbanco T, Elmore JG, Fitzgerald PS, Fossa A, Harcourt K, *et al.* Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes. JAMA Netw Open 2020;3:e205867. https://doi.org/10.1001/amanetworkopen.2020.5867.
104. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, *et al.* Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc JAMIA 2020;27:1999–2010. https://doi.org/10.1093/jamia/ocaa245.
105. 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;133:1–13. https://doi.org/10.1016/j.jclinepi.2020.12.019.
106. 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 JAMIA 2021;28:2251–7. https://doi.org/10.1093/jamia/ocab132.
107. Tahar K, Martin T, Mou Y, Verbuecheln R, Graessner H, Krefting D. Rare Diseases in Hospital Information Systems-An Interoperable Methodology for Distributed Data Quality Assessments. Methods Inf Med 2023;62:71–89. https://doi.org/10.1055/a-2006-1018.
108. Syed R, Eden R, Makasi T, Chukwudi I, Mamudu A, Kamalpour M, *et al.* Digital Health Data Quality Issues: Systematic Review. J Med Internet Res 2023;25:e42615. https://doi.org/10.2196/42615.
109. Lee S, Roh G-H, Kim J-Y, Ho Lee Y, Woo H, Lee S. Effective data quality management for electronic medical record data using SMART DATA. Int J Med Inf 2023;180:105262. https://doi.org/10.1016/j.ijmedinf.2023.105262.
110. Declerck J, Kalra D, Vander Stichele R, Coorevits P. Frameworks, Dimensions, Definitions of Aspects, and Assessment Methods for the Appraisal of Quality of Health Data for Secondary Use: Comprehensive Overview of Reviews. JMIR Med Inform 2024;12:e51560. https://doi.org/10.2196/51560.
111. Eder J, Gottweis H, Zatloukal K. IT solutions for privacy protection in biobanking. Public Health Genomics 2012;15:254–62. https://doi.org/10.1159/000336663.
112. Atreya RV, Smith JC, McCoy AB, Malin B, Miller RA. Reducing patient re-identification risk for laboratory results within research datasets. J Am Med Inform Assoc JAMIA 2013;20:95–101. https://doi.org/10.1136/amiajnl-2012-001026.
113. Kho AN, Cashy JP, Jackson KL, Pah AR, Goel S, Boehnke J, *et al.* Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Inform Assoc JAMIA 2015;22:1072–80. https://doi.org/10.1093/jamia/ocv038.
114. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat Methods 2016;13:251–6. https://doi.org/10.1038/nmeth.3746.
115. Prasser F, Spengler H, Bild R, Eicher J, Kuhn KA. Privacy-enhancing ETL-processes for biomedical data. Int J Med Inf 2019;126:72–81. https://doi.org/10.1016/j.ijmedinf.2019.03.006.
116. Zhang Z, Yan C, Mesa DA, Sun J, Malin BA. Ensuring electronic medical record simulation through better training, modeling, and

evaluation. J Am Med Inform Assoc JAMIA 2020;27:99–108. https://doi.org/10.1093/jamia/ocz161.

117. 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;19:255. https://doi.org/10.1186/s12911-019-0983-9.

118. Foraker RE, Yu SC, Gupta A, Michelson AP, Pineda Soto JA, Colvin R, *et al.* Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. JAMIA Open 2020;3:557–66. https://doi.org/10.1093/jamiaopen/ooaa060.

119. Marsolo K, Kiernan D, Toh S, Phua J, Louzao D, Haynes K, *et al.* Assessing the impact of privacy-preserving record linkage on record overlap and patient demographic and clinical characteristics in PCORnet®, the National Patient-Centered Clinical Research Network. J Am Med Inform Assoc JAMIA 2023;30:447–55. https://doi.org/10.1093/jamia/ocac229.

120. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc JAMIA 2013;20:117–21. https://doi.org/10.1136/amiajnl-2012-001145.

121. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, *et al.* Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc JAMIA 2013;20:e147-154. https://doi.org/10.1136/amiajnl-2012-000896.

122. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc JAMIA 2016;23:1046–52. https://doi.org/10.1093/jamia/ocv202.

123. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. J Biomed Inform 2015;58:156–65. https://doi.org/10.1016/j.jbi.2015.10.001.

124. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc JAMIA 2016;23:731–40. https://doi.org/10.1093/jamia/ocw011.

125. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, *et al.* Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc JAMIA 2016;23:1166–73. https://doi.org/10.1093/jamia/ocw028.

126. Luo L, Li L, Hu J, Wang X, Hou B, Zhang T, *et al.* A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. BMC Med Inform Decis Mak 2016;16:114. https://doi.org/10.1186/s12911-016-0357-5.

127. 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. https://doi.org/10.1016/j.ijmedinf.2018.01.007.

128. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018;361:k1479. https://doi.org/10.1136/bmj.k1479.

129. 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;14:3426–44. https://doi.org/10.1038/s41596-019-0227-6.

130. Kashyap M, Seneviratne M, Banda JM, Falconer T, Ryu B, Yoo S, *et al.* Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network. J Am Med Inform Assoc JAMIA 2020;27:877–83. https://doi.org/10.1093/jamia/ocaa032.

131. Eaneff S, Obermeyer Z, Butte AJ. The Case for Algorithmic Stewardship for Artificial Intelligence and Machine Learning Technologies. JAMA 2020;324:1397–8. https://doi.org/10.1001/jama.2020.9371.

132. Estiri H, Strasser ZH, Murphy SN. High-throughput phenotyping with temporal sequences. J Am Med Inform Assoc JAMIA 2021;28:772–81. https://doi.org/10.1093/jamia/ocaa288.

133. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, Young A, Jelovsek JE, O'Brien C, *et al.* A framework for the oversight and local deployment of safe and high-quality prediction models. J Am Med Inform Assoc JAMIA 2022;29:1631–6. https://doi.org/10.1093/jamia/ocac078.

134. Ahuja Y, Zou Y, Verma 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. https://doi.org/10.1016/j.jbi.2022.104190.

135. Keloth VK, Banda JM, Gurley M, Heider PM, Kennedy G, Liu H, *et al.* Representing and utilizing clinical textual data for real world studies: An OHDSI approach. J Biomed Inform 2023;142:104343. https://doi.org/10.1016/j.jbi.2023.104343.

136. Hamidi B, Flume PA, Simpson KN, Alekseyenko AV. Not all phenotypes are created equal: covariates of success in e-phenotype specification. J Am Med Inform Assoc JAMIA 2023;30:213–21. https://doi.org/10.1093/jamia/ocac157.

137. Brandt PS, Kho A, Luo Y, Pacheco JA, Walunas TL, Hakonarson H, *et al.* Characterizing variability of electronic health record-driven phenotype definitions. J Am Med Inform Assoc JAMIA 2023;30:427–37. https://doi.org/10.1093/jamia/ocac235.

138. Huser V, Cimino JJ. Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials. PloS One 2013;8:e68409. https://doi.org/10.1371/journal.pone.0068409.

139. Heneberg P. From Excessive Journal Self-Cites to Citation Stacking: Analysis of Journal Self-Citation Kinetics in Search for Journals, Which Boost Their Scientometric Indicators. PloS One 2016;11:e0153730. https://doi.org/10.1371/journal.pone.0153730.

140. Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, *et al.* What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res 2021;23:e22219. https://doi.org/10.2196/22219.

141. McIntosh LD, Juehne A, Vitale CRH, Liu X, Alcoser R, Lukas JC, *et al.* Repeat: a framework to assess empirical reproducibility in biomedical research. BMC Med Res Methodol 2017;17:143. https://doi.org/10.1186/s12874-017-0377-6.

142. Slager SL, Weir CR, Kim H, Mostafa J, Del Fiol G. Physicians' perception of alternative displays of clinical research evidence for clinical decision support - A study with case vignettes. J Biomed Inform 2017;71S:S53–9. https://doi.org/10.1016/j.jbi.2017.01.007.

143. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, *et al.* Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform 2012;45:763–71. https://doi.org/10.1016/j.jbi.2012.01.009.

144. O'Driscoll A, Daugelaite J, Sleator RD. "Big data", Hadoop and cloud computing in genomics. J Biomed Inform 2013;46:774–81. https://doi.org/10.1016/j.jbi.2013.07.001.

145. Sahoo SS, Jayapandian C, Garg G, Kaffashi F, Chung S, Bozorgi A, *et al.* Heart beats in the cloud: distributed analysis of electrophysiological "Big Data" using cloud computing for epilepsy clinical research. J Am Med Inform Assoc JAMIA 2014;21:263–71. https://doi.org/10.1136/amiajnl-2013-002156.

146. Bellazzi R, Dagliati A, Sacchi L, Segagni D. Big Data Technologies: New Opportunities for Diabetes Management. J Diabetes Sci Technol 2015;9:1119–25. https://doi.177/1932296815583505.

147. 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;25:792–804. https://doi.org/10.1038/s41591-019-0414-6.

148. 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;12:5757. https://doi.org/10.1038/s41467-021-26040-1.

149. Brat GA, Weber GM, Gehlenborg N, Avillach P, Palmer NP, Chiovato L, *et al.* International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med 2020;3:109. https://doi.org/10.1038/s41746-020-00308-0.

150. Kim SYH, Miller FG. Informed consent for pragmatic trials--the integrated consent model. N Engl J Med 2014;370:769–72. https://doi.org/10.1056/NEJMhle1312508.

151. Sugarman J, Califf RM. Ethics and regulatory complexities for pragmatic clinical trials. JAMA 2014;311:2381–2. https://doi.org/10.1001/jama.2014.4164.

152. Gray EA, Thorpe JH. Comparative effectiveness research and big data: balancing potential with legal and ethical considerations. J Comp Eff Res 2015;4:61–74. https://doi.org/10.2217/cer.14.51.

153. Kim H, Bell E, Kim J, Sitapati A, Ramsdell J, Farcas C, *et al.* iCONCUR: informed consent for

clinical data and bio-sample use for research. J Am Med Inform Assoc JAMIA 2017;24:380–7. https://doi.org/10.1093/jamia/ocw115.

154. Lea NC, Nicholls J, Fitzpatrick NK. Between Scylla and Charybdis: Charting the Wicked Problem of Reusing Health Data for Clinical Research Informatics. Yearb Med Inform 2018;27:170–6. https://doi.rg/10.1055/s-0038-1641219.

155. Chandler R, Brady KT, Jerome RN, Eder M, Rothwell E, Brownley KA, *et al.* Broad-scale informed consent: A survey of the CTSA landscape. J Clin Transl Sci 2019;3:253–60. https://doi.org/10.1017/cts.2019.397.

156. 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;22:51. https://doi.org/10.1186/s12910-021-00616-4.

157. Christofidou M, Lea N, Coorevits P. A Literature Review on the GDPR, COVID-19 and the Ethical Considerations of Data Protection During a Time of Crisis. Yearb Med Inform 2021;30:226–32. https://doi.rg/10.1055/s-0041-1726512.

158. Kargl M, Plass M, Müller H. A Literature Review on Ethics for AI in Biomedical Research and Biobanking. Yearb Med Inform 2022;31:152–60. https://doi.rg/10.1055/s-0042-1742516.

159. Larson DB, Harvey H, Rubin DL, Irani N, Tse JR, Langlotz CP. Regulatory Frameworks for Development and Evaluation of Artificial Intelligence-Based Diagnostic Imaging Algorithms: Summary and Recommendations. J Am Coll Radiol JACR 2021;18:413–24. https://doi.org/10.1016/j.jacr.2020.09.060.

160. Solomonides AE, Koski E, Atabaki SM, Weinberg S, McGreevey JD, Kannry JL, *et al.* Defining AMIA's artificial intelligence principles. J Am Med Inform Assoc JAMIA 2022;29:585–91. https://doi.org/10.1093/jamia/ocac006.

161. Ohno-Machado L. NIH's Big Data to Knowledge initiative and the advancement of biomedical informatics. J Am Med Inform Assoc JAMIA 2014;21:193. https://doi.org/10.1136/amiajnl-2014-002666.

162. Holmes JH, Elliott TE, Brown JS, Raebel MA, Davidson A, Nelson AF, *et al.* Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. J Am Med Inform Assoc JAMIA 2014;21:730–6. https://doi.org/10.1136/amiajnl-2013-002370.

163. Wilcox A, Randhawa G, Embi P, Cao H, Kuperman GJ. Sustainability considerations for health research and analytic data infrastructures. EGEMS Wash DC 2014;2:1113. https://doi.org/10.13063/2327-9214.1113.

164. Unertl KM, Schaefbauer CL, Campbell TR, Senteio C, Siek KA, Bakken S, *et al.* Integrating community-based participatory research and informatics approaches to improve the engagement and health of underserved populations. J Am Med Inform Assoc JAMIA 2016;23:60–73. https://doi.org/10.1093/jamia/ocv094.

165. Longo DL, Drazen JM. Data Sharing. N Engl J Med 2016;374:276–7. https://doi.org/10.1056/NEJMe1516564.

166. Wiley LK, Tarczy-Hornoch P, Denny JC, Freimuth RR, Overby CL, Shah N, *et al.* Harnessing next-generation informatics for personalizing medicine: a report from AMIA's 2014 Health Policy Invitational Meeting. J Am Med Inform Assoc JAMIA 2016;23:413–9. https://doi.org/10.1093/jamia/ocv111.

167. McCray AT, Glaser J, Koppel R, Langlotz CP, Silverstein J. Health IT vendors and the academic community: The 2014 ACMI debate. J Biomed Inform 2016;60:365–75. https://doi.org/10.1016/j.jbi.2016.03.003.

168. Spencer K, Sanders C, Whitley EA, Lund D, Kaye J, Dixon WG. Patient Perspectives on Sharing Anonymized Personal Health Data Using a Digital System for Dynamic Consent and Research Feedback: A Qualitative Study. J Med Internet Res 2016;18:e66. https://doi.org/10.2196/jmir.5011.

169. Valenta AL, Meagher EA, Tachinardi U, Starren J. Core informatics competencies for clinical and translational scientists: what do our customers and collaborators need to know? J Am Med Inform Assoc JAMIA 2016;23:835–9. https://doi.org/10.1093/jamia/ocw047.

170. Snyder CF, Smith KC, Bantug ET, Tolbert EE, Blackford AL, Brundage MD, *et al.* What do these scores mean? Presenting patient-reported outcomes data to patients and clinicians to improve interpretability. Cancer 2017;123:1848–59. https://doi.org/10.1002/cncr.30530.

171. Adler-Milstein J, Embi PJ, Middleton B, Sarkar IN, Smith J. Crossing the health IT chasm: considerations and policy recommendations to overcome current challenges and enable value-based care. J Am Med Inform Assoc JAMIA 2017;24:1036–43. https://doi.org/10.1093/jamia/ocx017.

172. Richesson RL, Green BB, Laws R, Puro J, Kahn MG, Bauck A, *et al.* Pragmatic (trial) informatics: a perspective from the NIH Health Care Systems Research Collaboratory. J Am Med Inform Assoc JAMIA 2017;24:996–1001. https://doi.org/10.1093/jamia/ocx016.

173. Sydes MR, Ashby D. Data Authorship as an Incentive to Data Sharing. N Engl J Med 2017;377:402. https://doi.org/10.1056/NEJMc1707245.

174. Sanchez-Pinto LN, Mosa ASM, Fultz-Hollis K, Tachinardi U, Barnett WK, Embi PJ. The Emerging Role of the Chief Research Informatics Officer in Academic Health Centers. Appl Clin Inform 2017;8:845–53. https://doi.org/10.4338/ACI-2017-04-RA-0062.

175. Franklin P, Chenok K, Lavalee D, Love R, Paxton L, Segal C, *et al.* Framework To Guide The Collection And Use Of Patient-Reported Outcome Measures In The Learning Healthcare System. EGEMS Wash DC 2017;5:17. https://doi.org/10.5334/egems.227.

176. 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;19:90. https://doi.org/10.1186/s12911-019-0799-7.

177. Kalra D, Stroetmann V, Sundgren M, Dupont D, Schlünder I, Thienpont G, *et al.* The European Institute for Innovation through Health Data. Learn Health Syst 2017;1:e10008. https://doi.org/10.1002/lrh2.10008.

178. Embi PJ, Richesson R, Tenenbaum J, Kannry J, Friedman C, Sarkar IN, *et al.* Reimagining the research-practice relationship: policy recommendations for informatics-enabled evidence-generation across the US health system. JAMIA Open 2019;2:2–9. https://doi.org/10.1093/jamiaopen/ooy056.

179. Zayas-Cabán T, Chaney KJ, Rucker DW. National health information technology priorities for research: A policy and development agenda. J Am Med Inform Assoc JAMIA 2020;27:652–7. https://doi.org/10.1093/jamia/ocaa008.

180. Byrne LM, Cook SK, Kennedy N, Russell M, Jerome RN, Tan J, *et al.* Opening doors to clinical trial participation among Hispanics: Lessons learned from the Spanish translation of ResearchMatch. J Clin Transl Sci 2020;5:e46. https://doi.org/10.1017/cts.2020.539.

181. Zayas-Cabán T, Abernethy AP, Brennan PF, Devaney S, Kerlavage AR, Ramoni R, *et al.* Leveraging the health information technology infrastructure to advance federal research priorities. J Am Med Inform Assoc JAMIA 2020;27:647–51. https://doi.org/10.1093/jamia/ocaa011.

182. 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;131:104096. https://doi.org/10.1016/j.jbi.2022.104096.

183. Peters AE, Ogunniyi MO, Hegde SM, Bianco C, Ghafghazi S, Hernandez AF, *et al.* A multicenter program for electronic health record screening for patients with heart failure with preserved ejection fraction: Lessons from the DELIVER-EHR initiative. Contemp Clin Trials 2022;121:106924. https://doi.org/10.1016/j.cct.2022.106924.

184. Maurud S, Henni SH, Moen A. Health Equity in Clinical Research Informatics. Yearb Med Inform 2023;32:138–45. https://doi.rg/10.1055/s-0043-1768720.

185. Alberto IRI, Alberto NRI, Ghosh AK, Jain B, Jayakumar S, Martinez-Martin N, *et al.* The impact of commercial health datasets on medical research and health-care algorithms. Lancet Digit Health 2023;5:e288–94. https://doi.org/10.1016/S2589-7500(23)00025-0.

186. Russell ES, Aubrun E, Moga DC, Guedes S, Camelo Castillo W, Hardy JR, *et al.* FDA draft guidance to improve clinical trial diversity: Opportunities for pharmacoepidemiology. J Clin Transl Sci 2023;7:e101. https://doi.org/10.1017/cts.2023.515.

187. Royce TJ, Zhao Y, Ryals CA. Improving Diversity in Clinical Trials by Using Real-world Data to Define Eligibility Criteria. JAMA Oncol 2023;9:455–6. https://doi.org/10.1001/jamaoncol.2022.7170.

188. Patel VN, Kaelber DC. Using aggregated, de-identified electronic health record data for multivariate pharmacosurveillance: a case study of azathioprine. J Biomed Inform 2014;52:36–42. https://doi.org/10.1016/j.jbi.2013.10.009.

189. Trifirò G, Coloma PM, Rijnbeek PR, Romio

S, Mosseveld B, Weibel D, *et al.* Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J Intern Med 2014;275:551–61. https://doi.org/10.1111/joim.12159.
190. Koutkias VG, Jaulent M-C. Computational approaches for pharmacovigilance signal detection: toward integrated and semantically-enriched frameworks. Drug Saf 2015;38:219–32. https://doi.org/10.1007/s40264-015-0278-8.
191. Marsolo K, Margolis PA, Forrest CB, Colletti RB, Hutton JJ. A Digital Architecture for a Network-Based Learning Health System: Integrating Chronic Care Management, Quality Improvement, and Research. EGEMS Wash DC 2015;3:1168. https://doi.org/10.13063/2327-9214.1168.
192. Harle CA, Lipori G, Hurley RW. Collecting, Integrating, and Disseminating Patient-Reported Outcomes for Research in a Learning Healthcare System. EGEMS Wash DC 2016;4:1240. https://doi.org/10.13063/2327-9214.1240.
193. 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 Lond Engl 2019;394:1816–26. https://doi.org/10.1016/S0140-6736(19)32317-7.
194. Platt JE, Raj M, Wienroth M. An Analysis of the Learning Health System in Its First Decade in Practice: Scoping Review. J Med Internet Res 2020;22:e17026. https://doi.org/10.2196/17026.
195. Geva A, Stedman JP, Manzi SF, Lin C, Savova GK, Avillach P, *et al.* Adverse drug event presentation and tracking (ADEPT): semiautomated, high throughput pharmacovigilance using real-world data. JAMIA Open 2020;3:413–21. https://doi.org/10.1093/jamiaopen/ooaa031.
196. Hartley DM, Seid M. Collaborative learning health systems: Science and practice. Learn Health Syst 2021;5:e10286. https://doi.org/10.1002/lrh2.10286.
197. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, *et al.* Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Am Med Inform Assoc JAMIA 2015;22:179–91. https://doi.org/10.1136/amiajnl-2014-002649.
198. Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open 2020;3:326–31. https://doi.org/10.1093/jamiaopen/ooaa033.
199. Mullen CG, Houlihan JY, Stroo M, Deeter CE, Freel SA, Padget AM, *et al.* Leveraging retooled clinical research infrastructure for Clinical Research Management System implementation at a large Academic Medical Center. J Clin Transl Sci 2023;7:e127. https://doi.org/10.1017/cts.2023.550.
200. Kwee SA, Wong LL, Ludema C, Deng CK, Taira D, Seto T, *et al.* Target Trial Emulation: A Design Tool for Cancer Clinical Trials. JCO Clin Cancer Inform 2023;7:e2200140. https://doi.org/10.1200/CCI.22.00140.
201. Wang SV, Schneeweiss S, RCT-DUPLICATE Initiative, Franklin JM, Desai RJ, Feldman W, *et al.* Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. JAMA 2023;329:1376–85. https://doi.org/10.1001/jama.2023.4221.
202. Ganzinger M, Blumenstock M, Fürstberger A, Greulich L, Kestler HA, Marschollek M, *et al.* Federated electronic data capture (fEDC): Architecture and prototype. J Biomed Inform 2023;138:104280. https://doi.org/10.1016/j.jbi.2023.104280.
203. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation Frameworks for Artificial Intelligence Translation Into Health Care Practice: Scoping Review. J Med Internet Res 2022;24:e32215. https://doi.org/10.2196/32215.
204. Crossnohere NL, Elsaid M, Paskett J, Bose-Brill S, Bridges JFP. Guidelines for Artificial Intelligence in Medicine: Literature Review and Content Analysis of Frameworks. J Med Internet Res 2022;24:e36823. https://doi.org/10.2196/36823.
205. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group, *et al.* Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351–63. https://doi.org/10.1038/s41591-020-1037-7.

Correspondence to:
Christel Daniel
Sorbonne Université,
INSERM UMR_S 1142, LIMICS, F-75006, Paris, France
Email: christel.daniel@aphp.fr

Copyright