Published online: 2020-08-21

Review of Clinical Research Informatics

Anthony Solomonides
Outcomes Research Network, Research Institute, NorthShore University HealthSystem, Evanston, IL, USA

Summary

Objectives: Clinical Research Informatics (CRI) declares its scope in its name, but its content, both in terms of the clinical research it supports—and sometimes initiates—and the methods it has developed over time, reach much further than the name suggests. The goal of this review is to celebrate the extraordinary diversity of activity and of results, not as a prize-giving pageant, but in recognition of the field, the community that both serves and is sustained by it, and of its interdisciplinarity and its international dimension.

Methods: Beyond personal awareness of a range of work commensurate with the author's own research, it is clear that, even with a thorough literature search, a comprehensive review is impossible. Moreover, the field has grown and subdivided to an extent that makes it very hard for one individual to be familiar with every branch or with more than a few branches in any depth. A literature survey was conducted that focused on informatics-related terms in the general biomedical and healthcare literature, and specific concerns ("artificial intelligence", "data models", "analytics", etc.) in the biomedical informatics (BMI) literature. In addition to a selection from the results from these searches, suggestive references within them were also considered.

Results: The substantive sections of the paper—Artificial Intelligence, Machine Learning, and "Big Data" Analytics; Common Data Models, Data Quality, and Standards; Phenotyping and Cohort Discovery; Privacy: Deidentification, Distributed Computation, Blockchain; Causal Inference and Real-World Evidence—provide broad coverage of these active research areas, with, no doubt, a bias towards this reviewer's interests and preferences, landing on a number of papers that stood out in one way or another, or, alternatively, exemplified a particular line of work.

Conclusions: CRI is thriving, not only in the familiar major centers of research, but more widely, throughout the world. This is not to pretend that the distribution is uniform, but to highlight the potential for this domain to play a prominent role in supporting progress in medicine, healthcare, and wellbeing everywhere. We conclude with the observation that CRI and its practitioners would make apt stewards of the new medical knowledge that their methods will bring forward.

Keywords

Clinical data, common data models, biomedical data quality, data analytics, artificial intelligence

Yearb Med Inform 2020:193-202
http://dx.doi.org/10.1055/s-0040-1701988

Introduction

The most comprehensive study of the field of Clinical Research Informatics (CRI), a study now ten years old, is that led by Drs. Embi and Payne [1]. Certainly, the landscape has changed in this time, but the basis for drawing the boundaries where they were drawn and for the subdivision of topics, methods, and tools, still provides a useful framework within which to assess more recent work. The study's definition touched on the demarcation of clinical research by the National Institutes of Health, but broke free to provide an independent formulation:

> *"Clinical Research Informatics involves the use of informatics in the discovery and management of new knowledge relating to health and disease. It includes management of information related to clinical trials and also involves informatics related to secondary research use of clinical data. Clinical research informatics and translational bioinformatics are the primary domains related to informatics activities to support translational research.*

This delineation of CRI has been elucidated further in occasional editorials and in Peter Embi's annual Year-in-Review keynote addresses to the American Medical Informatics Association (AMIA) Summit conferences, 2011 to 2018 [2]. To those who have attended these presentations, the format and breezy style of this paper could appear somewhat familiar, but we make no claim to similarity of method or of standard of commentary.

Many topics of interest today may also have featured ten years ago, but there are some that did not figure at all in 2009, and others whose centrality today might not have been predicted back then. Among these we may count Artificial Intelligence (AI), especially in the form of Machine/Deep Learning, our understanding of causal inference, the shifting trend in the use of "Real World Evidence" often gathered by networks using one or other of the several "common data models", to say nothing of innovations like deidentification and deduplication of overlapping populations using keyed cryptographic functions and—even more recently—experiments with blockchain. Big data analytics, visualization, descriptive and predictive applications were new then, and one readily recalls that next in the parade of adjectives they were expected to be "prescriptive"—which has indeed come to pass.

Methods

A broad search for research and review articles published in the years 2018 and 2019 was conducted with search terms "Clinical Research" and "Informatics" in general medical journals. More narrowly specified searches in biomedical informatics journals on terms "artificial intelligence", "phenotyping", "cohort discovery", "real world evidence", "causal inference", "natural language", "cognition", "diagnosis", and combinations, were supplemented with references from the most relevant publications from such searches. Certain topics, notably AI and issues around privacy, have also attracted a great deal of comment; a sample of viewpoints and opinion pieces has been included.

The aim throughout, however imperfectly realized, was to be open to research from anywhere in the world, so long as it advanced the field of CRI. A secondary aim was to allow for some historical flow, so that if a trend or line of work began, say, two years ago, it would not automatically be excluded on grounds of age. Among the many subtopics worthy of review, we have focused on the highly active areas of AI, on approaches to clinical trials including the use of real world evidence, on the broad area

of cohort discovery, de-identification and privacy protection, including the possible value of blockchain in this respect, and on language and cognition in CRI. The final list of papers from which the selection presented here was made was classified into categories as tabulated in Table 1.

It must be freely admitted that the selection reflects the personal tastes and preferences of the author, and should not be regarded as in any sense some kind of prize shortlist. That the author is proud to be associated with the field of CRI should be abundantly clear. That the selection of publications is also limited by the extent of his familiarity with the field is also necessarily true.

Table 1 Results of the literature search for CRI papers.

PRINCIPAL REVIEW CATEGORIES	COUNT
AI (MACHINE-LEARNING, IMAGING, AND NATURAL-LANGUAGE-PROCESSING)	99
ANALYTICS - DATA SCIENCE	13
BLOCKCHAIN	12
CAUSAL INFERENCE	33
COGNITION AND DIAGNOSIS	10
DATA, DATA QUALITY, AND DATA QUERIES	59
DE-IDENTIFICATION AND DISTRIBUTED PARADIGMS	42
ETHICAL, LEGAL, AND SOCIAL ISSUES	23
ECONOMICS, BUSINESS, AND IMPLEMENTATION	19
PHENOTYPING AND COHORT DISCOVERY	67
REAL WORLD EVIDENCE	39
MISCELLANEOUS	24
TOTAL	440

Results

Artificial Intelligence, Machine Learning, and "Big Data" Analytics

Artificial Intelligence has risen to prominence in a way that belies its years of overpromise and underdelivery. What has brought this about is, in part, slow maturation, but the transformation arguably began with the apparent abandonment of logic as the foundation of AI (think of "expert systems") in preference for a plurality of data and approaches to "learning", *i.e.,* development of models that fit, describe, and may ultimately explain a set of facts or observations, in the sense of providing a means to comprehend the pattern of the data, not merely the individual data points. Indeed, not all individual data points need to be explainable in this way. One speaker mused in a 2009 keynote [3] whether this trend meant "the abandonment of soundness for completeness", alluding to a well-known "incompleteness" theorem in logic (any sound formal system, complex enough to support arithmetic, necessarily includes assertions that are true but not provable within the rules of the system itself. A sufficiently rich complete system, therefore, cannot be sound). It is in this context that we speak of "comprehending" a pattern in the data even when some data defy the pattern. A thorough exposition of Deep Learning has been provided by the pioneers, Le Cun, Bengio, and Hinton [4]. A somewhat less technical review of its potential and clear discussion of its role in augmenting, rather than supplanting, human intelligence has been provided by Rajkomar, Dean, and Kohane [5].

The breadth of papers on AI in medicine and healthcare certainly defies easy summary. Indeed, any selection is likely to be representative of the tastes of the reviewer: this limitation must be admitted. Among papers that show the immense promise of Machine Learning in healthcare is an interesting analysis of the potential of Electronic Health Records (EHRs) to yield useful knowledge by Rajkomar *et al.* [6], appropriately published in a journal that is itself newly dedicated to the field of Digital Medicine. These authors adopted a deep learning approach using the entire EHR and addressed four representative questions using a single data structure: for outcomes, risk of death; for quality, risk of readmission; for resource efficiency, length of stay; and for the semantic value of the record, patient diagnoses. Their approach avoids the variable selection problem and outperforms across various indices. However, it is a retrospective study, so the challenge remains to build predictive models through time in the EHR and validate through prospective studies. Other notable work in the field takes a radically different approach. A team at University of California San Francisco (UCSF) reports on a project with similar predictive goals for patients with one particular condition, rheumatoid arthritis (RA) [7]. In this study, variables were selected based on known clinical significance, though not necessarily known to have predictive value. A strict phenotype for RA was applied in two diverse settings, a university hospital and a safety net hospital. Encouragingly, the results were applicable in both, suggesting that robust models may be transposable to new settings once developed. Yet another approach, taking its cue from process mining, addresses the broad problem of diagnostic error for undifferentiated chief complaints [8]. How is the sequence of events following presentation with abdominal pain best understood and visualized? Are some diagnostic trajectories more effective than others? How do time and timing impact the process? How well does this approach translate from one chief complaint to another—say from abdominal pain to dizziness? Many such questions remain to be addressed. Discovery or refinement of disease phenotypes is another potential application of AI. In the case of sepsis, it has

been observed that improved understanding of the immune response has not translated into improved treatments. This is partly due to the enormous range of clinical and biological features that figure in the definition of the syndrome. A team at the University of Pittsburgh reports on a study [9] that identified four new clinical phenotypes that may help explain diverse treatment effects and can guide the design of clinical trials and future treatment regimens.

AI, and its promise, limitations, and implications, have attracted voluminous commentary from experts and from anticipated beneficiaries. The Journal of the American Medical Association has paid close attention to such questions. We note in particular a guide to reading the literature [10], an accompanying editorial [11], and a viewpoint review [12] of the National Academy of Medicine's comprehensive exploration of AI in healthcare [13]. Possible biases in the design and development of AI systems in conjunction with EHRs have also been explored [14], as has their remediation [15] and the potential legal liability risk for a provider using AI [16]. Considering the influential regulatory framework in the US on Software as a Medical Device, how should the lifecycle of an AI system be viewed, especially if it is adaptive and—at least in theory—self-improving [17]? The "black box" paradigm is an apt description for much modern AI. Models are constructed and decisions made with virtually no explanation. This is in stark contrast to "classical" AI which was formal logic-based and could in the main provide a logical audit trail for a decision. The desirability for explanation in general has been recognized and begun to be addressed in Explainable AI (XAI) [18, 19]. In healthcare, given all the other concerns reviewed here, the need for explanation goes beyond desirable to essential. Work in this area is underway, at this stage mainly in the engineering domain [20], but with applicability to healthcare already under consideration [21].

A number of viewpoints and opinion pieces have addressed ethical, legal, and social issues. Is it possible for an AI tool to monitor the status of a mental health patient [22]? Would a conversational agent—"agent" being a term of art for software that can initiate real world actions and, in many cases, act autonomously—be an appropriate tool to address underserved mental health needs [23]? How does AI mediate or interfere in the relationship between physician and patient [24]? Conversely, what is its potential to reduce provider burden and burnout [25]?

A somewhat contrasting approach [26] that leans more heavily on statistical methods [27] is variously described as "Data Science", "Big Data", or "Analytics," although its practitioners sometimes describe it as "AI" [28]. It has been successful in improving clinical operations, delivery of care, and health system administration. The goal is often to target a particular performance index (*e.g.,* average length of stay, 30-day readmission, or immunization rate) or a status index for the patient population (e.g., percentage of controlled patients with diabetes, or of those with asthma who experience exacerbations). A typical technique is to identify patients at risk and devise targeted interventions. Poor data quality sometimes impairs the predictive power of these methods [29]. It is generally considered most advantageous to implement models in the EHR so that triggers can fire alerts for action [30]. It is the ambition of Learning Health Systems [31, 32] to have cumulative evidence from practice improve the predictive value of the models even as they are being used; naturally, this raises some of the "black box" issues alluded to above. Finally, we note that analytics has also served patients in activities bordering on Citizen Science [33].

Common Data Models, Data Quality, and Standards

EHR systems are optimized for transactions, so that providers will experience minimal delays in their interactions with patients. Data generated in this way is subsequently stored in a database "normalized" so as to minimize duplication of information (thus, risk of inconsistency), yet at the same time amenable to search by means of a structured query language. Clinical research often revolves around the discovery of particular, sometimes very complex, cohorts of patients. Common Data Models, as they have come to be known, provide a further filter for the organization and storage of data in a highly standardized form, so that even data from different institutions may be navigated using the same basic queries. Among the most popular such models, i2b2 ("Informatics for Integrating Biology and the Bedside") and OMOP ("Observational Medical Outcomes Partnership") were created in 2005 and 2008, respectively, both essentially with observational data in mind, for clinical research in the case of i2b2 and to study effects of interventions and drugs for OMOP. Motivated by regulatory changes, the Federal Drug Administration's (FDA) Mini-Sentinel post-marketing drug surveillance program also created a common data model, and this provided the inspiration for (and largely lent its design to) the first issue of Patient-Centered Outcomes Research Network's Common Data Model (PCORnet CDM). These have opened up a number of avenues for research based on "real world data"(RWD)—data collected in the course of healthcare delivery or even from the use of health-related applications on mobile devices. Research on the scope and validity of RWD and the ways in which the analysis of RWD may lead to "real world evidence" (RWE) deserve a section of its own, but it is worth alluding here to the FDA's definition and discussion of these terms [34]: "Real-world data are the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. … Real-world evidence is the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from the analysis of RWD."

An example of the breadth of rich data and study potential in an environment of independent entities using a common data model may be seen in the study Short- and Long-Term Effects of Antibiotics on Childhood Growth. Using the strict criteria of same day height and weight in each of three distinct age periods (0 to <12, 12 to <30, and 30 to <72 months), working across 35 institutions, a diverse cohort of 362,550 children were found to be eligible for the study [35]. Of these, just over 58% had received at least one antibiotic prescription, with over 33% receiving a broad-spectrum antibiotic. The cohort was large enough to allow for adjustment for complex chronic conditions.

In children without such a condition, the odds ratio for overweight or obesity was 1.05 (CI 1.03 to 1.09) for those with at least one antibiotic before age 24 months. The effect was thus shown to be real, but small [36]. The study group was able to identify 53,320 mother–child pairs to consider whether antibiotic use by mothers had an effect on childhood weight; taking into account timing during pregnancy, dose-response, spectrum and class of antibiotics, the study found no associations between maternal antibiotic use and the distribution of BMI at age 5 [37]. With an eye to what matters to stakeholders—parents and primary physicians—the study also considered whether these findings would influence prescribing patterns and parental expectations; the answer was unambiguously "no" [38]. They were also able to examine data quality by comparing prescriptions and dispensing in 200,395 records and identified gaps in these data, although prescription data were adequate for the question at hand [39]. Finally, in a technical proof of principle, they showed that a form of distributed regression analysis, avoiding the aggregation of patient-level data, generated results comparable to those of the main study [40].

What goes into providing data for such a study? The process begins with the creation of common data model-conforming data marts or mappings to enterprise or research data warehouses; this is termed "extract-transform-load" (ETL). In research, this must be followed by "phenotyping", translating the inclusion and exclusion criteria and the defined information for the desired cohort in the form of a query that captures as precisely as possible the required data. Intervening between the two steps in this ideal sequence of operations (or perhaps part of a good ETL process) is data quality analysis. Each of these stages presents certain problems and attracts attention from researchers in the effort to bridge gaps. Notwithstanding the popularity of common data models such as the PCORnet CDM and OMOP in the US (and increasingly elsewhere), a good deal of research still addresses the choice of clinical data model and interoperability between such models. The HL7 FHIR interface standard is increasingly accepted as a way forward.

Looking at particular examples of work in this area, a group in Germany has set out to model the ETL process [41] and along the way define quality checks [42] and provenance standards [43]. This is an interesting way to marry process principles and implementation at the level of technology. They link their provenance work in particular both to technical and to administrative or regulatory requirements, so that researchers would not have to engage in separate activities in using data for research and in ensuring that it is handled according to all legal and ethical requirements. Yet another contribution from Germany [44] is preoccupied with the completeness and syntactic accuracy of data from a heterogeneous network of institutions, using a (logically) central metadata repository as its reference point. An Australian contribution [45] rooted in business systems seeks a design method, or at least a set of design principles, towards a unified view of data quality management in healthcare environments, alongside methods and tools derived from that design view.

In the United States, much of the attention to ETL processes and data quality has centered on the major common data models. Collaborations around both the PCORnet CDM and OMOP have focused on data quality, with PCORnet requiring quarterly "data characterization" or "hygiene" queries [46] and major tool developments by the OHDSI (Observational Health Data Sciences and Informatics [47]) collaborative. A number of notable efforts have thus cumulatively created an impressive collection of results. Using a data quality (DQ) ontology of their own devising in 2017 [48], a group led by Michael Kahn analyzed and mapped DQ approaches in six networks [49]. In the same year, Weiskopf *et al.,* published a guideline for DQ assessment [50]. In 2018, there followed a contribution by Gold *et al.,* on the challenges of concept value sets and their possible reuse; indicative of the acuity of the challenge is that not all co-authors could sign up to every view expressed in the paper [51]! Rogers *et al.,* [52] then analyzed data element–function combinations in checks from two environments, ultimately identifying 751 unique elements and 24 unique functions, supporting their systematic approach to DQ check definition. Most recently, Chunhua Weng has offered a lifecycle perspective, indeed a philosophy, for clinical DQ for research [53], while Seneviratne, Kahn, and Hernandez-Boussard have provided an overview of challenges for the merging of heterogeneous data sets, with an eye both on integration across institutions (where adherence to standards may be sufficient) and across "modalities", the latter term interpreted in its widest possible sense, encompassing genomics, imaging, and patient-reported data from wearables [54].

It is necessary to add two more observations to this section. One concerns the Fast Health Interoperability Resources (FHIR) standard specification and the other the commercially supported data marts that have made a significant mark on institutions. The relative proliferation of data models, and the passionate attachment of each one's proponents to the primacy of their chosen model, have resulted in a great deal of duplication of work—the very avoidance of which was one of the drivers for their introduction in the first place. It is debated whether one model or another should be taken as the definitive basis for research data in an institution and how other data needs would then be met. Thrown into this mix, FHIR, an interface or data exchange standard, has at times been spoken of as a "meta-model" from which all others can be derived. However, to quote an authority on this question, "FHIR's purpose is not to define a persistence layer—it's to define a data exchange layer. It exists to define data structures used to pass information from one system to another. That doesn't mean that you can't use FHIR models to define how you store data, only that FHIR isn't designed for that purpose and doesn't provide any guidance about how to do that" [55]. FHIR is thus likely to be an excellent approach for defining data elements to extract from data close to care delivery; how that data is then stored and manipulated for research remains the question, so the choice of model remains fraught.

The second observation is that in the midst of this babel of models a number of academically well-informed companies have proposed a different, private business model—varying somewhat by company—whereby the data is curated and made available for anonymous search, often aggregated by geo-

graphic area or nationally. Institutions need not be identified unless they wish to be, and the cost, which may be considerable, is borne by the commercial clients of these systems, most obviously pharmaceutical companies, eager to define cohorts for trials, to gauge how much of a market there might be for a drug, and so on. At present, these systems co-exist with home-grown or academically developed public systems (cf. *Leaf* [56] for an excellent example), but there is a sense in which they are in competition, so this is a space to be watched.

Phenotyping and Cohort Discovery

These observations on data models, and their realization as real repositories—data marts or data warehouses—naturally lead to the question of their use. There are possible uses in quality improvement and in public health, but our main focus is clinical research. Here, then, is the place to acknowledge and celebrate the successes of major international consortia and other loose affiliations that have created banks of phenotypes and cohort discovery tools to help navigate the data in standard models. By "phenotype" we mean the criteria that identify patients with a given condition or disease. These may be complex: *e.g.,* patients with type II diabetes (DM2) may be identified by having a pertinent ICD9 or ICD10 code in their problem list or elsewhere in their record, or may be on a combination of therapies that is uniquely appropriate for DM2, or may be suffering from a complication, such as maculopathy, that has been annotated to indicate that it is due to DM2. Although translational informatics is often interpreted in a genomic context, it should be acknowledged that the need to match phenotypes to genotypes has provided considerable stimulus to phenotyping from EHRs.

Work begun by the Electronic Medical Records and Genomics (eMERGE) [57] network and the NIH Collaboratory was first reported in 2016 [58] and has continued to grow in the most sustained and focused effort, both to generate precise phenotypes and to sharpen existing ones. The concept has indeed been accepted more widely and applied to good effect. A report [59] in a nephrology journal on phenotyping pediatric glomerular disease, a rare condition, is accompanied by an appreciative editorial [60] recognizing similar efforts in the discipline. Looking to a particularly difficult case, Koola *et al.,* demonstrate improved phenotyping of hepatorenal syndrome, a difficult sub-phenotype of acute kidney injury [61]. Pacheco *et al.,* use methods from the Phenotype Execution Modeling Architecture (PhEMA) project [62] to demonstrate the portability of a benign prostatic hyperplasia phenotype across a number of institutions [63]. Taylor *et al.,* use developed phenotypes to identify patterns of comorbidities in eMERGE network institutions [64]. That this work is far from easy is readily demonstrated by the difficulties described in other works, such as the study by Fawcett *et al.,* in the UK [65] and that by Ando *et al.,* in Japan [66]. Nevertheless, independent efforts to phenotype particular conditions still arise and show considerable promise. We note in hematology the work of Singh *et al.*, [67] and in psychiatric genomics that of Smoller [68]. Increasing reliance on EHR phenotyping is reflected not only in the proliferation of papers applying one or other approach to particular, often complex, conditions, but also in notable dissemination efforts, including an extended exposition by Pendergrass and Crawford aimed at human geneticists [69].

Alternative methodologies also appear in the literature. A semantic approach by Zhang *et al.,* takes its cue from difficulties encountered in translating specifications (*e.g.,* in PheKB) to query code and to specialize a phenotype to each instance of a data repository [70]. Reflecting the transition we have observed in AI applications, Banda *et al.,* outline a possible trajectory from "rule-based" phenotyping to machine learning models and suggest a research program to complete the move [71]. Among ML approaches, Ding *et al.,* adopt "multitask learning", and find that multitask deep learning nets outperform the simpler single-task nets—a counterintuitive observation. Yet another interesting study combined ML and rule-based approaches to identify entities and relations, providing a natural language interface to clinical databases [72].

Clinical trial recruitment is often the driving reason for phenotyping. Several papers with a focus on recruitment came to this reviewer's attention. A clinical trial recruitment planning framework by the Clinical Trials Transformation Initiative offers evidence-based recommendations on trial design, trial feasibility, and communication [73]. A Veterans' Affairs team describes a holistic approach to clinical trial management, including identification of possible subjects and recruitment, based on its Cooperative Studies Program working together with VA Informatics [74]. An oncology team at Vanderbilt and Rush reports on an ambitious framework capturing multiple aspects of clinical trial management, including evaluation [75]. Finally, a team from Seoul, Korea, reviews the creation, deployment, and evaluation of an entire Clinical Trial Management System which offers the full range of functions required for recruitment and full ethical and regulatory compliance [76].

We will conclude this section with a brief mention of a reflexive analysis of the work that goes into the creation of a library of phenotypes [77]. A retrospective analysis of phenotyping algorithms in the eMERGE network identified nearly 500 "clauses" (phenotype criteria) associated with over 1100 tasks, some 60% of which are related to knowledge and interpretation and 40% to programming. In each case, portability (of knowledge, of interpretation, and of programming) was graded on a scale of 0–3, resulting in each phenotype receiving a score that reflects expert perception of its portability. Having commended this work, a parallel reading recommendation should be the analysis of patients' and clinical professionals' "data work" by Fiske, Prainsack, and Buyx [78]. Here the ambiguity of "data" ("givens") and their contextual dependency are discussed with insight and with empathy for all participants.

Privacy: Deidentification, Distributed Computation, Blockchain

The last two decades in biomedical informatics have seen enormous growth in large-scale collaborations and attempts to combine and share data to gain power in results, to achieve greater diversity, and to provide the much vaunted "evidence" necessary for evidence-based practice. The rationale and the challenges are well described in Haynes *et al.*,

[79]. One of the most frequently encountered obstacles to data sharing for research is the concern over patient privacy—and rightly so, of course. In most jurisdictions, there is some legal or regulatory protection for personal health information. The acronym "PHI", for Protected Health Information, is precisely defined in US legislation, but also serves as a shorthand for what may commonly be considered personal and private health information. A critical aspect that is derived from the relevant US act (the Health Information Portability and Accountability Act, HIPAA) is that any value or code that is derived from PHI is itself PHI, unless a certain kind of one-way cryptographic "hash" function ("hashing" for short) is used in the derivation. Data that has been thus transformed is often termed "de-identified", although experts now are careful to circumscribe claims of de-identification. Significant reasons for this are the availability of other data sets, either as public goods or otherwise available for purchase, and the possible use of similar methods to link individuals in one data set to those in the supposedly de-identified collection. Cross correlation of information thus obtained can lead to reidentification of at least some individuals in the list. Where it can be done securely, an additional benefit of hashing is the possibility of deduplication of patients that attend more than one healthcare system, thus enabling a more nearly complete picture of a person's record to be aggregated without identifying the patient by name. Following significant advances in the last few years [80, 81], more recent work has extended and exploited these methods [82, 83]. Kayaalp has surveyed a number of approaches [84].

A particularly interesting method has been advanced by Hejblum *et al.,* in Boston [85]. It is clear that sufficient clinical details (such as diagnoses with encounter dates) may be enough to identify a subject uniquely, or very nearly so, in two coherent data sets. What if there are discrepancies between certain data elements in the two sets? The method presented allows them to compute the probability of identity even when certain elements do not agree. Other methods of de-duplication, not necessarily with anonymity, include a Bayesian method [86] adapted from astronomy—galaxies in different astronomical databases may not have matched names, but do have matched characteristics, by and large. Funded by the German Medical Informatics Initiative [87], the SMITH consortium is developing the infrastructure to support a network of Data Integration Centres (DIC), which will share services and functionality to provide access to the local hospitals' Electronic Medical Records (EMRs). Regulatory protections will be provided by data trustees and privacy management services, but DIC staff will be able to curate and amend EMR data in a core Health Data Storage. Secure multi-party computation features in a number of studies, including a persuasive two-party instance using garbled circuits in a geographically wide-ranging collaboration in the United States [88] and an Estonian report of a multi-party system based on the Sharemind platform that is "ready for practical use" [89]. Another technical aspect, which has received some attention, is the high combinatorial cost of pairwise comparison for de-duplication. An approach known as "blocking and windowing", which originates in the founding studies in statistical de-duplication, is used to reduce the dimensionality of the comparison space, and is still being refined in various ways [90]. Further methods of interest use Bloom filter pairs. The method of Brown *et al.,* exhibits good error tolerance [91], while that of Ranbaduge and Christen [92] includes the temporal information in records in its hashing process; this Australian contribution is all the more interesting in the light of extensive national data linking guidelines by the government [93].

Blockchain has been suggested as a possible answer to the challenges of anonymous data sharing. Indeed, in the world of Health IT, as one encounters it in practice in health systems, blockchain is being viewed with interest [94, 95]. Researchers have begun some exploratory work, but blockchain has not yet had wide adoption in the field. Some interesting work can be reported here. The Journal of Medical Systems has a special collection of papers on blockchain, including a study of a blockchain-based privacy-preserving record linkage (PPRL) solution [96]. Other studies of blockchain include a Swiss-American systematic review of oncology applications [97], 16 in all at the time of the study, distributed among countries, unsurprisingly USA (4 studies), Switzerland (2 studies), and Germany, Iraq, Taiwan, Italy, China (1 study per country), and a proof of principle study using a novel framework, HealthChain [98].

A radically different concept in privacy preserving analysis is distributed computation. A University of Pennsylvania-led team describes two performant algorithms [99, 100], differentiated by resource requirements and performance, which analyze data behind their home firewalls and aggregate statistical results, mainly regression models. Their particular success lies in controlling for data source heterogeneity and maintaining high faithfulness to gold standard (*i.e.,* analysis of aggregated data). On the technical front, another anticipated development is that of trustworthy databases as described by Rogers *et al.,* in which complementary scenarios of trusted database/untrustworthy analyst and untrustworthy cloud/trusted analyst motivate the technical requirements [101].

Much of the public concern with data sharing in healthcare [102] revolves around the known or alleged abuses that "big tech" stands accused of [103–105] and near certainty that deidentification [106], even when expertly done and certified, does not eliminate the risk of reidentification [107 – 109]. A contentious re-identification exercise was reported and commented on in the Journal of the American Medical Association (JAMA) in 2018 [110, 111]. As has been repeatedly pointed out by patient advocates, in a jurisdiction in which the possibility that certain kinds of health or long term care insurance, employment prospects, and other rights, may be limited by what is known about one's health status, privacy of PHI must be fiercely guarded.

Causal Inference and Real-World Evidence

A significant change in the regulatory environment impacted the world of food and drug law in mid-2017, although this was written up later in 2018:

> *"In June 2017, FDA approved a new indication for a medical device without requiring any new clinical trials. This approval marked the onset of a new era in drug and medical device regulation: the systemic use of "Real World Evidence"*

(RWE). FDA based its approval on records of the product's actual patient use rather than on randomized clinical trials" [112].

In some respects, biomedical research, informatics in particular, has been ahead of the game. In the wake of the realization that the gold standard for research—randomized clinical trials—is slow to deliver the hoped-for results and improvements, informatics has embraced observational data and outcomes research. Among the most insightful lines of inquiry focuses on the question, if we want observational studies to deliver comparably robust results to those of clinical trials, how should we design our observational studies? There is progress to report on a number of fronts, including data collection, observational trial design, and causal analysis.

Specialty journals as well as informatics titles have been reporting on particular efforts to collect data for research in the process of delivery of care. Examples include clinical oncology [113], neurology [114], nutrition and endocrinology [115, 116], and pharmacovigilance [117] to name but a few, not that any of these are complacent or make an easy equation between RWE and real-world data (RWD). It is generally appreciated that turning RWD into RWE requires work—often ingenious and complex work [118 – 121].

This reviewer's enthusiasm for Hernán's and his collaborators' work will be obvious sooner or later, so I will leap right in. Their "second chance" paper [122] lays out the tasks ahead with great clarity and analytic perspicacity. Koch's postulates in microbiology are now obsolescent, but if one were to aspire, for observational studies, to the degree of rigor they implied, this might be a good place to start. Also worth following is a spirited defense of causality and debate between Hernán and several other scientists with pro and con views in the American Journal of Public Health [123]. Another debate on causality also took place in the Journal od the American Medical Informatics Association (JAMIA) and casts a different light on the question [124]. Two books are likely to prove highly influential in this domain: Pearl and Mackenzie's The Book of Why [125], which provides the intellectual framework for a science of causality, and a forthcoming textbook by Hernán and Robbins, based on courses delivered at Harvard [126].

Conclusion

The goal of this review was not so much to select the most spectacular work in CRI, though much of the work discussed is indeed superlative, but to convey a sense of how much is going on in the field, how widely spread it is, and how fruitful it is proving, even without touching on the favored science or our times—genomics. Having said that, it is also the case that a good deal of CRI is undertaken in conjunction with genomics and, more broadly, translational science. This is surely all to the good and bodes well for those who choose to work closer to the clinical field. The more our evidence base expands and deepens, and the clearer the interaction of social determinants and biology becomes, the more likely it is that the task of knowledge management in the field will fall to CRI experts.

Conflict of Interest

Anthony Solomonides is a co-author of reference [8], offered as an exemplar of process mining work, and a collaborating co-author of references [35] and [36], part of an extended series of papers on the study Short- and Long-Term Effects of Antibiotics on Childhood Growth.

Acknowledgements

I had access to two libraries with complementary e-collections, that of my current institution, NorthShore University HealthSystem, and that of my former university, UWE, Bristol, UK. I am grateful to Ms. Linda Feinberg, MLIS, Library Director at NorthShore, for her assistance in literature search. I also owe thanks to the leadership group of the American Medical Informatics Associations CRI Working Group, whose board members, notably Monika Ahuja, Kate Fultz Hollis, Abu Mosa, Nelson Sánchez-Pinto, Lincoln Sheets, Ana Szarfman, Chunhua Weng, and Tamara Winden suggested papers for review.

References

1. Embi PJ, Payne PRO. Clinical Research Informatics: Challenges, Opportunities and Definition for an Emerging Domain. J Am Med Inform Assoc 2009;16:316–27.
2. Embi P. AMIA CRI Years in Review. Presentations available at http://www.embi.net/cri-years-in-review.html
3. van Harmelen F. Protégé 2009. Invited Talk. Presentation available at: https://protege.stanford.edu/conference/2009/slides/FrankvanHarmelen_ProtegeConf09.pdf
4. Le Cun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
5. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019;380:1347-58.
6. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Medi 2018;1:18.
7. Norgeot B, Glicksberg BS, Trupin L, Lituiev D, Gianfrancesco M, Oskotsky B, et al. Assessment of a Deep Learning Model Based on Electronic Health Record Data to Forecast Clinical Outcomes in Patients With Rheumatoid Arthritis. JAMA Netw Open 2019;2(3):e190606.
8. Rao G, Kirley K, Epner P, Zhang Y, Bauer V, Padman R, et al. Identifying, Analyzing, and Visualizing Diagnostic Paths for Patients with Nonspecific Abdominal Pain. Appl Clin Inform 2018;9:905–13.
9. Seymour CW, Kennedy JN, Wang S, Chang CCH, Elliot CF, Xu Z, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. JAMA 2019;321(20):2003-17.
10. Liu Y, Chen PHC, Krause J, Peng L. How to Read Articles That Use Machine Learning. Users' Guides to the Medical Literature. JAMA 2019;322(18):1806-16.
11. Doshi-Velez F, Perlis RH. Evaluating Machine Learning Articles. JAMA 2019;322(18):1777-9.
12. Matheny ME, Whicher D, Israni ST. Artificial Intelligence in Health Care- A Report From the National Academy of Medicine. JAMA 2019, Dec 17. Online ahead of print.
13. Matheny M, Israni ST, Ahmed M, Whicher D, editors. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. NAM Special Publication. Washington, DC: National Academy of Medicine; 2019. Available at: https://nam.edu/artificial-intelligence-special-publication/
14. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med 2018;178(11):1544-7.
15. Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. JAMA 2019;322(24):2377-8.
16. Nicholson Price W II, Gerke S, Cohen IG. Potential Liability for Physicians Using Artificial Intelligence. JAMA 2019;322(18):1765-6.
17. Hwang TJ, Kesselheim AS, Vokinger KN. Lifecycle Regulation of Artificial Intelligence– and Machine Learning–Based Software Devices in

Medicine. JAMA 2019;322(23):2285-6.
18. Samek W, Wiegand T, *Müller* K-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ITU Journal: ICT Discoveries 2017; Special Issue 1. International Telecommunication Union; 2018.
19. Samek W, Montavon G, Vedaldi A, Hansen LK, Klaus-Robert Müller K-R, editors. Explainable AI- Interpreting, Explaining and Visualizing Deep Learning LNAI 11700. Springer; 2019.
20. Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access October 2018.
21. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): towards Medical XAI. https://arxiv.org/abs/1907.07374
22. Insel TR. Digital Phenotyping- Technology for a New Science of Behavior. JAMA 2017;318(13):1215-6.
23. Miner AS, Milstein A, Hancock JT. Talking to Machines About Personal Mental Health Problems. JAMA 2017;318(13):1217-8.
24. Nundy S, Montgomery T, Wachter RM. Promoting Trust Between Patients and Physicians in the Era of Artificial Intelligence. JAMA 2019;322(6):497-8.
25. Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a Physician- Humanism and Artificial Intelligence. JAMA 2018;319(1):19-20.
26. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12–22.
27. Liu VX, Bates DW, Wiens J, Shah NH. The number needed to benefit: estimating the value of predictive analytics in healthcare. J Am Med Inform Assoc 2019;26(12):1655–9 .
28. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc 2019;26(12):1651–4.
29. Newcomer SR, Xu S, Kulldorff M, Daley MF, Fireman B, Glanz JM. A primer on quantitative bias analysis with positive predictive values in research using electronic health data. J Am Med Inform Assoc 2019; 26(12):1664–74.
30. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 2018 Aug 1;25(8):969-75.
31. Institute of Medicine (US) Roundtable on Evidence-Based Medicine; Olsen L, Aisner D, McGinnis JM, editors. The Learning Healthcare System: Workshop Summary. Washington (DC): National Academies Press (US); 2007. (This was the first of a number of Workshop Reports issued by the Roundtable on Evidence-Based Medicine. See also 32 below.)
32. Institute of Medicine. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. Washington, DC: The National Academies Press; 2011. https://doi.org/10.17226/12912
33. Evans BJ, Krumholz HM. People-powered data collaboratives: fueling data science with the health-related experiences of individuals. J Am Med Inform Assoc 2019;26(2):159–61.
34. Federal Drug Administration. Real World Evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
35. Block JP, Bailey LC, Gillman MW, Lunsford D, Boone-Heinonen J, Cleveland LP, et al. PCORnet Antibiotics and Childhood Growth Study: Process for Cohort Creation and Cohort Description. Acad Pediatr 2018;18(5):569 – 76.
36. Block JP, Bailey LC, Gillman MW, Lunsford D, Daley MF, Eneli I, et al. Early Antibiotic Exposure and Weight Outcomes in Young Children. Pediatrics 2018;142(6):e20180290 .
37. Heerman WJ, Daley MF, Boone-Heinonen J, Rifas-Shiman SL, Bailey CL, Forrest CB, et al. Maternal antibiotic use during pregnancy and childhood obesity at age 5 years. Int J Obes (Lond) 2019;43:1202–9.
38. Lipstein EA, Block JP, Dodds C, Forrest CB, Heerman WJ, Law JK, et al. Early Antibiotics and Childhood Obesity: Do Future Risks Matter to Parents and Physicians? Clin Pediatr (Phila) 2019;58(2):191–8.
39. Lin PD, Daley MF, Boone-Heinonen J, Rifas-Shiman SL, Bailey CL, Forrest CB, et al; PCORnet Antibiotics and Childhood Growth Study Group. Comparing Prescribing and Dispensing Data of the PCORnet Common Data Model Within PCORnet Antibiotics and Childhood Growth Study. EGEMS (Wash DC) 2019;7(1):11.
40. Toh S, Rifas-Shiman SL, Lin PD, Bailey CL, Forrest CB, Horgan CE, et al. Privacy-protecting multivariable-adjusted distributed regression analysis for multi-center pediatric study. Pediatr Res 2020;87(6):1086-92.
41. Tute E, Steiner J. Modeling of ETL-Processes and Processed Information in Clinical Data Warehousing. Stud Health Technol Inform 2018;248:204-11.
42. Tute E, Wulff A, Marschollek M, Gietzelt M. Clinical Information Model Based Data Quality Checks: Theory and Example. Stud Health Technol Inform 2019;258:80-4.
43. Parciak M, Bauer C, Bender T, Lodahl R, Schreiweis B, Tute E, et al. Provenance Solutions for Medical Research in Heterogeneous IT-Infrastructure: An Implementation Roadmap. Stud Health Technol Inform 2019; 264:298 – 302.
44. Juárez D, Schmidt EE, Stahl-Toyota S, Ückert F, Lablan M. A Generic Method and Implementation to Evaluate and Improve Data Quality in Distributed Research Networks. Methods Inf Med 2019;58(2-03):86-93.
45. Bai L, Meredith R, Burstein F. A data quality framework, method and tools for managing data quality in a health care setting: an action case study. Journal of Decision Systems 2018;27(Suppl 1):144-54.
46. Ong T, Pradhananga R, Holve E, Kahn MG. A Framework for Classification of Electronic Health Data Extraction-Transformation-Loading Challenges in Data Network Participation. EGEMS 2017;5(1):10.
47. https://www.ohdsi.org/
48. 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(1):1244.
49. Callahan TJ, Bauck AE, Bertoch D, Brown J, Khare R, Ryan PB, et al. A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. EGEMS (Wash DC) 2017;5(1):8.
50. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. EGEMS (Wash DC) 2017;5(1):14.
51. Gold S, Batch A, McClure R, Jiang G, Kharrazi H, Saripalle R, et al. Clinical Concept Value Sets and Interoperability in Health Data Analytics. AMIA Annu Symp Proc 2018;2018:480–9.
52. Rogers JR, Callahan TJ, Kang T, Bauck A, Khare R, Brown JS, et al. A Data Element-Function Conceptual Model for Data Quality Checks. EGEMS (Wash DC) 2019;7(1):17.
53. Weng C. Clinical data quality: a data life cycle perspective. Biostat Epidemiol 2019;4(1):6-14.
54. Seneviratne MG, Kahn MG, Hernandez-Boussard T. Merging heterogeneous clinical data to enable knowledge discovery. Pac Symp Biocomput 2019;24:439–43.
55. FHIR & Relational Data Mode (http://community.fhir.org/t/fhir-relational-data-model/427) Response by Lloyd McKenzie to a question on the suitability of FHIR to serve as a relational model.
56. Dobbins NJ, Spital CH, Black RA, Morrison JM, de Veer B, Zampino E, et al. Leaf: an open-source, model-agnostic, data-driven web application for cohort discovery and translational biomedical research. J Am Med Inform Assoc 2020 Jan 1;27(1):109-18.
57. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013;15(10):761–71.
58. 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 2016 Nov;23(6):1046-52.
59. Denburg MR, Razzaghi H, Bailey LC, Soranno DE, Pollack AH, Dharnidharka VR, et al. Using Electronic Health Record Data to Rapidly Identify Children with Glomerular Disease for Clinical Research. J Am Soc Nephrol 2019 Dec;30(12):2427-35.
60. Glenn D, Gibson KL. Finding That Needle in the Haystack: Computable Phenotypes. J Am Soc Nephrol 2019 Dec;30(12):2279-80.
61. Koola JD, Davis SE, Al-Nimri O, Parr SK, Fabbri D, Malin BA, et al. Development of an automated phenotyping algorithm for hepatorenal syndrome. J Biomed Inform 2018;80:87–95.
62. Phenotype Execution Modeling Architecture (PhEMA) http://informatics.mayo.edu/phema/index.php/Main_Page
63. Pacheco JA, Rasmussen LV, Kiefer RC, Campion TR, Speltz P, Carroll RJ, et al. A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions

and electronic health record environments. J Am Med Inform Assoc 2018;25(11):1540–6.
64. Taylor CO, Lemke KW, Richards TM, Roe KD, He T, Arruda-Olson A, et al. Comorbidity Characterization Among eMERGE Institutions: A Pilot Evaluation with the Johns Hopkins Adjusted Clinical Groups® System. AMIA Jt Summits Transl Sci Proc 2019;2019:145–52.
65. Fawcett N, Young B, Peto L, Quan TP, Gillott R, Wu J, et al. 'Caveat emptor': the cautionary tale of endocarditis and the potential pitfalls of clinical coding data-an electronic health records study. BMC Med 2019;17(1):169.
66. Ando T, Ooba N, Mochizuki M, Koide D, Kimura K, Lee SL, et al. Positive predictive value of ICD-10 codes for acute myocardial infarction in Japan: a validation study at a single center. BMC Health Serv Res 2018;18(1):895.
67. Singh A, Mora J, Panepinto JA. Identification of patients with hemoglobin SS/Sβ0 thalassemia disease and pain crises within electronic health records. Blood Adv 2018;2(11):1172–9.
68. Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. Am J Med Genet B Neuropsychiatr Genet 2018;177(7):601–12.
69. Pendergrass SA, Crawford DC. Using Electronic Health Records To Generate Phenotypes For Research. Curr Protoc Hum Genet 2019;100(1):e80.
70. Zhang H, He Z, He X, Guo Y, Nelson DR, Modave F, et al. Computable Eligibility Criteria through Ontology-driven Data Access: A Case Study of Hepatitis C Virus Trials. AMIA Annu Symp Proc 2018;2018:1601–10.
71. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci 2018 Jul;1:53-68.
72. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc 2019;26(4):294–305.
73. Huang GD, Bull J, Johnston McKee K, Mahon E, Harper B, Roberts JN; CTTI Recruitment Project Team. Clinical trials recruitment planning: A proposed framework from the Clinical Trials Transformation Initiative. Contemp Clin Trials 2018;66:74–9.
74. Velarde KE, Romesser JM, Johnson MR, Clegg DO, Efimova O, Oostema SJ, et al. An initiative using informatics to facilitate clinical research planning and recruitment in the VA health care system. Contemp Clin Trials Commun 2018;11:107–12.
75. Jain NM, Culley A, Knoop T, Micheel C, Osterman T, Levy M. Conceptual Framework to Support Clinical Trial Optimization and End-to-End Enrollment Workflow. JCO Clin Cancer Inform 2019;3:1–10.
76. Park YR, Yoon YJ, Koo H, Yoo S, Choi C-M, Beck S-H, et al. Utilization of a Clinical Trial Management System for the Whole Clinical Trial Process as an Integrated Database: System Development. J Med Internet Res 2018;20(4):e103.
77. Shang N, Liu C, Rasmussen LV, Ta CN, Caroll RJ, Benoit B, et al. Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network. J Biomed Inform 2019 Nov;99:103293.
78. Fiske A, Prainsack B, Buyx A. Data Work: Meaning-Making in the Era of Data-Rich Medicine. J Med Internet Res 2019 Jul 9;21(7):e11672.
79. Haynes K, Selvam N, Cziraky MJ. Bidirectional Data Collaborations in Distributed Research. EGEMS (Wash DC) 2016;4(2):1205.
80. 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 2015;22(5):1072–80.
81. Bennett TD, Dean JM, Keenan HT, McGlincy MH, Thomas AM, Cook LJ. Linked Records of Children with Traumatic Brain Injury. Probabilistic Linkage without Use of Protected Health Information. Methods Inf Med 2015;54(4):328–37.
82. Zimmerman LP, Goel S, Sathar S, Gladfelter CE, Onate A, Kane LL, et al. A Novel Patient Recruitment Strategy: Patient Selection Directly from the Community through Linkage to Clinical Data. Appl Clin Inform 2018;9(1):114–21.
83. CAPriCORN Chicago Homelessness Project (See https://www.capricorncdrn.org/projects/homelessness-project/)
84. Kayaalp M. Patient Privacy in the Era of Big Data. Balkan Med J 2018;35(1):8–17.
85. Hejblum BP, Weber GM, Liao KP, Palmer NP, Churchill S, Shadick NA, et al. Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. Sci Data 2019;6:180298.
86. Hurley PD, Oliver S, Mehta A. Creating longitudinal datasets and cleaning existing data identifiers in a cystic fibrosis registry using a novel Bayesian probabilistic approach from astronomy. PLoS One 2018;13(7):e0199815.
87. Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V, et al. Smart Medical Information Technology for Healthcare (SMITH). Methods Inf Med 2018;57(S 01):e92–e105.
88. Chen F, Jiang X, Wang S, Schilling LM, Meeker D, Ong T, et al. Perfectly Secure and Efficient Two-Party Electronic-Health-Record Linkage. IEEE Internet Comput 2018;22(2):32–41. doi:10.1109/MIC.2018.112102542
89. Laud P, Pankova A. Privacy-preserving record linkage in large databases using secure multiparty computation. BMC Med Genomics 2018;11(Suppl 4):84.
90. Sohail A, Yousaf MM. A proficient cost reduction framework for de-duplication of records in data integration. BMC Med Inform Decis Mak 2016;16:42.
91. Brown AP, Randall SM, Ferrante AM, Semmens JB, Boyd JH. Estimating parameters for probabilistic linkage of privacy-preserved datasets. BMC Med Res Methodol 2017;17(1):95.
92. Ranbaduge T, Christen P. A scalable privacy-preserving framework for temporal record linkage. Knowl Inf Syst 2020;62:45-78.
93. Australian Government. Open Data Toolkit: Data Linking (Updated August 2018; accessed December 2019.) https://toolkit.data.gov.au/Data_Linking_Information_Series_Contents_page.html
94. Mackey TK, Kuo T-T, Gummadi B, Clauson KA, Church G, Grishin D, et al. 'Fit-for-purpose?' - Challenges and opportunities for applications of blockchain technology in the future of healthcare. BMC Med 2019;17(1):68.
95. Hussien HM, Yasin SM, Udzir SNI, Zaidan AA, Zaidan BB. A Systematic Review for Enabling of Develop a Blockchain Technology in Healthcare Application: Taxonomy, Substantially Analysis, Motivations, Challenges, Recommendations and Future Direction. J Med Syst 2019;43(10):320.
96. Zhang A, Lin X. Towards Secure and Privacy-Preserving Data Sharing in e-Health Systems via Consortium Blockchain. J Med Syst 2018;42:140.
97. Dubovitskaya A, Novotny P, Xu Z, Wang F. Applications of Blockchain Technology for Data-Sharing in Oncology: Results from a Systematic Literature Review. Oncology 2019 Dec 3;1-9.
98. Hylock RH, Zeng X. A Blockchain Framework for Patient-Centered Health Records and Exchange (HealthChain): Evaluation and Proof-of-Concept Study. J Med Internet Res 2019;21(8):e13592.
99. Duan R, Boland MR, Liu Z, Liu Y, Chang HH, Xu H, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. J Am Med Inform Assoc 2020;27(3):376-85.
100. Tong J, Duan R, Li R, Scheuemie MJ, Moore JH, Chen Y. Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data. Pac Symp Biocomput 2020;25:695-706.
101. Rogers J, Bater J, He X, Machanavajjhala A, Suresh M, Wang X. Privacy Changes Everything . VLDB Poly'19 Towards Polystores that manage multiple Databases, Privacy, Security and/or Policy Issues for Heterogenous Data. VLDB'19, the 45th International Conference on Very Large Data Bases, Los Angeles, California - August 26-30, 2019. Available at: users.eecs.northwestern.edu/~jennie/pubs/privacy-changes-everything.pdf
102. Copeland R, Mattioli D, Evans M. Paging Dr. Google: How the Tech Giant Is Laying Claim to Health Data. Wall Street Journal, January 11, 2020. https://www.wsj.com/articles/paging-dr-google-how-the-tech-giant-is-laying-claim-to-health-data-11578719700. (Includes Dr. David Feinberg's first interview as VP for Google Health.)
103. Zuboff S. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: Public Affairs; 2019.
104. Information Commissioner's Office. Royal Free - Google DeepMind trial failed to comply with data protection law. July 2017. Available at: https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/.
105. Cohen JK. Google, UChicago Med sued over data-sharing practices. Modern Healthcare, June 27, 2019. https://www.modernhealthcare.com/node/952326/printable/print.

Cohen JK. Google, Ascension data partnership sparks federal probe. Modern Healthcare, November 13, 2019. https://www.modernhealthcare.com/node/959776/printable/print.
106. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. J Med Internet Res 2019;21(5):e13484.
107. Yoo J, Thaler A, Sweeney L, Zang J. Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data. Technology Science. 2018100901. October 09, 2018. https://techscience.org/a/2018100901
108. Janmey V, Elkin PL. Re-Identification Risk in HIPAA De-Identified Datasets: The MVA Attack. AMIA Annu Symp Proc 2018;2018:1329–37.
109. Simon GE, Shortreed SM, Coley RY, Penfold RB, Rossom RC, Waitzfelder BE, et al. Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records. EGEMS (Wash DC) 2019;7(1):6.
110. Na L, Yang C, Lo C, Zhao F, Fukuoka Y, Aswani A. Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning. JAMA Netw Open 2018;1(8):e186040.
111. McCoy TH, Hughes MC. Preserving Patient Confidentiality as Data Grow: Implications of the Ability to Reidentify Physical Activity Data. JAMA Netw Open 2018;1(8):e186029.
112. Belson NA. FDA's Historical Use of "Real World Evidence". Food and Drug Law Institute, Update Magazine, August/September 2018. https://www.fdli.org/2018/08/update-fdas-historical-use-of-real-world-evidence/
See also in the same issue:
Onwudiwe NC, Tenenbaum K, Boise BH, Elton J, Manning M. Real World Evidence: Implications and Challenges for Medical Product Communications in an Evolving Regulatory Landscape. Food and Drug Law Institute, Update Magazine, August/September 2018. https://www.fdli.org/2018/08/update-real-world-evidence-implications-and-challenges-for-medical-product-communications-in-an-evolving-regulatory-landscape/
113. Law JH, Pettengell C, Le LW, Aviv S, DeMarco P, Merritt DC, et al. Generating real-world evidence: Using automated data extraction to replace manual chart review. J Clin Oncol 2019;37:15: e18096.
114. Zhang R, Simon G, Yu F. Advancing Alzheimer's research: A review of big data promises. Int J Med Inform 2017;106:48–56.
115. He Z, Rizvi RF, Yang F, Adam TJ, Zhang R. Comparing the Study Populations in Dietary Supplement and Drug Clinical Trials for Metabolic Syndrome and Related Disorders. AMIA Jt Summits Transl Sci Proc 2019;2019:799–808.
116. Patorno E, Gopalakrishnan C, Franklin JM, Brodovicz KG, Masso-Gonzalez E, Bartels DB, et al. Claims-based studies of oral glucose-lowering medications can achieve balance in critical clinical variables only observed in electronic health records. Diabetes Obes Metab 2018;20(4):974–84.
117. Sobel RE, Bate A, Reynold, .F. Real World Evidence: Time for a Switch? Drug Saf 2018;41:1309–12.
118. 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(27):7329–36.
119. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 2018;25(8):969–75.
120. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. Philos Trans A Math Phys Eng Sci 2018;376(2128):20170356.
121. von Lucadou M, Ganslandt T, Prokosch HU, Toddenroth D. Feasibility analysis of conducting observational studies with the electronic health record. BMC Med Inform Decis Mak 2019;19(1):202.
122. Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. CHANCE 2019;32(1):42-9.
123. Hernán MA. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. Am J Public Health 2018;108(5):616–9.
Begg MD, March D. Cause and Association: Missing the Forest for the Trees. Am J Public Health 2018;108(5):620.
Chiolero A. Data Are Not Enough-Hurray For Causality! Am J Public Health 2018;108(5):622.
Glymour MM, Hamad R. Causal Thinking as a Critical Tool for Eliminating Social Inequalities in Health. Am J Public Health 2018;108(5):623.
Jones HE, Schooling CM. Let's Require the "T-Word". Am J Public Health 2018;108(5):624.
Hernán M. The C-Word: The More We Discuss It, the Less Dirty It Sounds. Am J Public Health 2018;108(5):625–6.
124. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless…, J Am Med Inform Assoc 2019;26(12):1645–50.
Sperrin M, Jenkins D, Martin GP, Peek N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. J Am Med Inform Assoc 2019;26(12):1675–6.
Lenert MC, Matheny ME, Walsh CG. Explicit causal reasoning is preferred, but not necessary for pragmatic value. J Am Med Inform Assoc 2019;26(12)1677–8.
125. Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect. New York: Basic Books; 2018.
126. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC. Forthcoming 2020. (A preliminary version is available online at: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2019/11/ci_hernanrobins_10nov19.pdf).

Correspondence to:
Anthony Solomonides
Research Institute
NorthShore University HealthSystem
1001 University Place
Evanston, IL 60201
USA
E-mail: asolomonides@northshore.org;
tony.solomonides@gmail.com