Published in final edited form as:

Methods Inf Med. 2021 May ; 60(1-02): 32–48. doi:10.1055/s-0041-1731784.

Why Is the Electronic Health Record So Challenging for Research and Clinical Care?

John H. Holmes[1], **James Beinlich**[2], **Mary R. Boland**[1], **Kathryn H. Bowles**[3], **Yong Chen**[1], **Tessa S. Cook**[4], **George Demiris**[3], **Michael Draugelis**[5], **Laura Fluharty**[6], **Peter E. Gabriel**[4], **Robert Grundmeier**[7], **C. William Hanson**[8], **Daniel S. Herman**[9], **Blanca E. Himes**[1], **Rebecca A. Hubbard**[1], **Charles E. Kahn Jr.**[4], **Dokyoon Kim**[1], **Ross Koppel**[10], **Qi Long**[1], **Nebojsa Mirkovic**[11], **Jeffrey S. Morris**[1], **Danielle L. Mowery**[1], **Marylyn D. Ritchie**[12], **Ryan Urbanowicz**[1], **Jason H. Moore**[1]

[1]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States

[2]Information Technology Entity Services and Corporate Information Services, University of Pennsylvania Health System, Philadelphia, Pennsylvania, United States

[3]Department of Biobehavioral Health Sciences, University of Pennsylvania School of Nursing, Philadelphia, Pennsylvania, United States

[4]Department of Radiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States

[5]Department of Predictive Health Care, University of Pennsylvania Health System, Philadelphia, Pennsylvania, United States

[6]Clinical Research Operations, University of Pennsylvania Health System, Philadelphia, Pennsylvania, United States

[7]Department of Pediatrics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States

[8]Department of Anesthesiology and Critical Care, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States

[9]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine Philadelphia, Pennsylvania, United States

[10]Department of Sociology, University of Pennsylvania, Philadelphia, Pennsylvania, United States

[11]Department of Research Analytics, University of Pennsylvania Health System, Philadelphia, Pennsylvania, United States

Address for correspondence John H. Holmes, PhD, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, United States (jhholmes@pennmedicine.upenn.edu).

Conflict of Interest

T.S.C. reports grants from ACRIN, NIH, ACR, and RSNA, as well as royalties from the Osler Institute for lectures in 2013, outside the submitted work. D.S.H. reports grants and nonfinancial support from Roche Diagnostics, outside the submitted work. R.A.H. reports grants from Johnson & Johnson, Merck, and Pfizer, outside the submitted work. Q.L. reports grants from NIH, during the conduct of the study; grants from Pfizer and Bayer, outside the submitted work.

[12]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States

Abstract

Background—The electronic health record (EHR) has become increasingly ubiquitous. At the same time, health professionals have been turning to this resource for access to data that is needed for the delivery of health care and for clinical research. There is little doubt that the EHR has made both of these functions easier than earlier days when we relied on paper-based clinical records. Coupled with modern database and data warehouse systems, high-speed networks, and the ability to share clinical data with others are large number of challenges that arguably limit the optimal use of the EHR

Objectives—Our goal was to provide an exhaustive reference for those who use the EHR in clinical and research contexts, but also for health information systems professionals as they design, implement, and maintain EHR systems.

Methods—This study includes a panel of 24 biomedical informatics researchers, information technology professionals, and clinicians, all of whom have extensive experience in design, implementation, and maintenance of EHR systems, or in using the EHR as clinicians or researchers. All members of the panel are affiliated with Penn Medicine at the University of Pennsylvania and have experience with a variety of different EHR platforms and systems and how they have evolved over time.

Results—Each of the authors has shared their knowledge and experience in using the EHR in a suite of 20 short essays, each representing a specific challenge and classified according to a functional hierarchy of interlocking facets such as usability and usefulness, data quality, standards, governance, data integration, clinical care, and clinical research.

Conclusion—We provide here a set of perspectives on the challenges posed by the EHR to clinical and research users.

Keywords

electronic health records; user-computer interface; standards; medical informatics; systems integration

Introduction

The use of electronic health records (EHRs) has become much more pervasive over the past 5 years, to the extent that the vast majority of hospitals and office-based physicians in the United States have adopted an EHR system.[1,2] Concomitantly, there is considerable evidence in the literature and even from personal anecdotal reports of which many are aware that suggest difficulties in using the EHR for patient care and research. For example, Artis et al found that incomplete data can lead to misdiagnosis and medical error when clinicians rely on the EHR when rounding[3] and during hand-offs.[4] These findings were substantiated in another study which found that physicians frequently identified poor EHR usability as a barrier to finding patient information[5]. Aside from the patient care context, another study found that primary care practitioners encountered difficulty in generating quality

improvement reports from EHR systems and subsequently adhering to meaningful use care coordination criteria.[6] As a reaction to these difficulties, practitioners have developed workarounds to address unintended consequences of EHR use that could affect patient safety.[7,8] However, little work has been reported on the consequences of using workarounds but the available evidence suggests that EHRs are sufficiently inadequate to accommodate every clinical or research contingency.[9–13] Finally, the EHR is front and center in the fight against the novel coronavirus disease 2019 (COVID-19) pandemic. The need has never been greater for robust EHR systems that provide timely access to accurate and complete health care data by clinicians, public health agencies and professionals, biomedical informaticians, and researchers who are engaged in patient care, clinical characterization of COVID-19, and establishing distributed networks for large-scale pandemic surveillance. As a result of all of these concerns, we feel that it is important to review the reasons for why the EHR presents so many challenges for users in clinical and research domains.

Methods

Conceptual Framework

We have structured this paper according to a functional classification that was created by the authors to reflect challenges relating to our central theme which is the use of the EHR in clinical and research domains. Based on our experience in clinical care and research uses and deployments of EHR systems, we reached consensus on a series of activities associated with characteristics of EHRs that relate to the central theme. These activities and characteristics are represented as classification facets in our framework:

- EHR usability and usefulness: reflects the user experience during use of an EHR system, with implications for patient safety, ability to use the EHR effectively for clinical care, and research.
- EHR data quality: reflects the quality and reliability of data as used in clinical care and research.
- EHR standards: reflects adherence by the EHR to standards (vocabularies, ontologies, nomenclatures, and communication) and the effect of standards on EHR use.
- EHR governance: reflects policies and procedures enacted to ensure the proper use of EHR systems and data.
- EHR and other data integration: reflects the procedures and challenges posed by integration of EHR data with other data sources, including other EHR systems.

Within each facet are one or more subfacets that represent concepts that were identified by the authors to be of critical importance to clinical and clinical research informaticians, as well as practicing clinicians. It is important to note that these facets often interlock with each other, such that they can overlap and that their relationships are of critical importance in addressing the central question of the user of the EHR. The relationships between these facets are illustrated in this directed acyclic graph (Fig. 1).

Another way to visualize our conceptual framework is in the following Venn diagram (Fig. 2) where the overlap between the five domains is clearly illustrated, as is the central concept, use of the EHR in clinical care and research, which underpins the entire model.

We explore each of these facets in more detail throughout this paper. Rather than approach these as a scoping or systematic review, we felt it was important to share our views and experience as informatics professionals.

The Authors

Our panel of biomedical informatics researchers, information technology professionals, and clinicians describe an extensive list of characteristics of the EHR that make its design, implementation, and use difficult in clinical and research contexts. To accomplish this, collectively, we bring a combined 510 years of experience in working with EHRs in research and clinical contexts, and from a wide spectrum of perspectives. We represent all of the five domains commonly considered to comprise biomedical informatics, from translational to clinical to clinical research to consumer health to public health informatics. Each of us has had an active role in the design, implementation, use, or evaluation of EHR systems. In addition, each of us has an active role in the overall organization of this paper, while leveraging our specific expertise to focus on the sections of the paper that are relevant to that expertise. Our particular specialties and experience are provided in the Supplementary Appendix (available in the online version). As a result, we hope to bring our experience and expertise to bear in exploring each of the five facets that we feel are most important in answering the question: “Why is the electronic health record so challenging for research and clinical care?.”

Organization of the Paper

We offer a consideration on the organization of this paper. While it is organized according to the facets and subfacets in our conceptual framework that are relevant to each facet, we propose that clinical care and clinical research are inextricable at this time; the centrality of EHR systems and data in the service of clinical research bears out this observation and, we believe, justifies our organization. Where appropriate, we identify possible overlaps between various sections in the matrix presented in “Overlap of the Subfacets” section. We hope that approaching an answer to our question in this way does not confuse the reader.

Usability and Usefulness

Human Factors and Human Computer Interaction: How Do We Make It Easy for Researchers and Clinicians to Interact with Software and Computational Technology? (George Demiris)

Usability challenges with EHRs have been well documented; these challenges pertain to clinicians’ accessing and navigating these systems, retrieving and analyzing relevant information, and having to engage in redundant procedures or workarounds to complete tasks. Some of these same challenges are also present for researchers who rely on using EHRs to extract and process data. As an informatics discipline, human factors or ergonomics research focuses on the study of human computer interaction, the user experience related

to interface design, information processing, and overall perception of a system. These are critically important aspects of EHRs and have significant influence whether a system will be adopted and how effectively it will be used once introduced into the workflow. In the context of health care, human factors issues do not only affect the overall user experience but may actually impede efficiency of clinical operations through increased documentation time and disrupted information flow, and even patient safety (e.g., causing documentation errors or alert fatigue). In a systematic review of usability studies with EHRs, Zahabi et al identified 50 studies that highlighted usability issues such as violations of natural dialog, control consistency, effective use of language, and effective information presentation.[14] Furthermore, this review showcased how usability principles such as customization options, error prevention, minimization of cognitive load, and feedback are in many instances ignored or violated. Similarly, Roman et al. conducted a systematic review of usability literature examining navigation in EHRs and found that navigation actions (e.g., scrolling through a medication list) were frequently linked to specific violations of usability heuristic principles such as recognition rather than recall, flexibility and efficiency of use, and error prevention.[15]

The challenge with EHRs is that they often have been designed without a recognition of the complex and everchanging cognitive, collaborative, organizational, and structural aspects of interdisciplinary health care delivery. Such systems have traditionally been conceptualized through a billing or overall administrative lens which may not fully align with clinicians' and patients' perspectives or information needs. As a 2005 Institute of Medicine Report pointed out, "usability in software-intensive systems cannot be achieved by patching user friendly interfaces onto user-hostile system architectures."[16] To facilitate access to software and computation technology for researchers and clinicians, further work is needed in the following three distinct areas: (1) assessment of user information needs and preferences; (2) user training; and (3) expansion of a usability standards framework and certification for health information systems. While system designers often assume what clinicians and researchers may need when it comes to using a health information system, these actual end users are often not included in early design phases, or only a limited subset of users is invited to provide feedback in spite of the broad spectrum of distinct stakeholder needs that need to be addressed by the system. End users need to be engaged in all phases of design, implementation, and formative evaluation of systems. Similarly, user training and support resources need to be developed, recognizing the importance of timely assistance and the burden that traditional training sessions may introduce. Finally, both federal agencies and the industry have recognized usability and human–computer interaction issues as imperative to the success of these systems and have introduced criteria and guidelines for usability. For example, the Office of the National Coordinator (ONC) of Health Information Technology requires EHR vendors to employ a user-centered design process as part of its certification requirements.[17] More specifically, a vendor must attest to having utilized a user-centered design approach and provide summative usability testing findings on eight functions of the system.[17] However, this requirement alone does not seem sufficient to address challenges in usability and the growing dissatisfaction of clinicians with these systems; more extensive and purposeful summative testing with larger number and more inclusive demographics of participants has been suggested.[18] An expansion of the usability framework that informs

certification efforts and potentially additional certification requirements have the potential to better address these challenges.

Electronic Health Records and Inflexibility (Ross Koppel)

Many clinicians and analysts find EHRs inflexible. EHRs often do not allow the nuances needed to reflect the patient's reality. The categories may make little sense—too broad, demanding levels of granularity not available from the presenting patients, even from patient's laboratory data, or not allowing clinicians to reflect the ambiguity and messiness of real medical practice. Sometimes EHRs demand inputs that are simply absurd, for example, a forced field requiring the value of the differential between the left and right iris responses to light, even though there are millions of people with only one functioning eye. Very often, EHRs' inflexibility results from the expectations of the multiple masters that create or maintain them: insurance company or other reimbursement needs; regulatory bodies; local, state, and federal data demands; superannuated diagnosis-related groups (DRGs), the International Classification of Diseases, version 10 (ICD-10) codes, conflicts among services or professions, medical associations in the way they categorize or scale issues; new treatment protocols; new drugs or diseases not yet incorporated into the software; or demands for data unknown to local clinicians, such as full medication lists, patients' histories, allergies, countries visited, and others, all with no way for the clinician to explain or code the contradictions or data unavailability. Last, there is another source of EHR inflexibility that emerges from rules that prohibit post hoc updating earlier entries, e.g., later test results, better diagnoses, revised treatment plans or medications.

Data Quality

Bias and Fairness (Ryan Urbanowicz and Qi Long)

The concepts of bias and fairness are interlinked in biomedical research. Bias, defined as prejudice against a person or group, has many potential sources in EHR analyses.[19] Data may be collected from unequal demographics of a population and systemic bias can impact the exposures and treatment decisions of individual patients. Such biases can impact the fairness of statistical or machine learning (ML) modeling such that resulting prediction models may have dissimilar error rates across population subgroups or lead to unfavorable treatment decisions and then worse outcomes for marginalized subgroups defined by, namely, race, gender, or age.[20] For example, data from predominantly Caucasian samples in the Framingham Heart Study were used to predict cardiovascular events, but applications of these models to nonwhite populations yielded over- and underestimations of risk.[21] Another example is that race has been used as part of an algorithm for estimating kidney function in practice, and it has been argued that this may unduly restrict access to care for under-served minority groups.[22] Ultimately biases existing in the data will tend to be reflected by the statistical or ML strategies applied to model and make predictions in that same domain. Fairness in EHR analyses can be promoted by first identifying and then accounting for biases, for example, over- or undersampling, adjusting the weights of samples, as well as by adding fairness-aware constraints, to the model training process.[23]

Variability in Laboratory Data (Daniel Herman)

Variability in clinical practice and documentation makes it extremely challenging to derive generalizable knowledge and models from EHR data. To illustrate this, let us focus on one of the most standardized domains of clinical practice: clinical laboratory testing. The fundamental question is if you go to two different clinical practices and get the same blood test, will the results be equivalent? There is considerable regulation of laboratories and their tests aimed at minimizing the variability of such results.[24] Unfortunately, even when this system works perfectly, results for some tests can differ widely across laboratories. It depends on how well standardized and on how accurate the tests are. For tests where the targeted biological analyte is precisely defined and the methods are mature, such as studies of the concentration of sodium ions in blood, test results tend to be very consistent across laboratories.[25] But, for more complex macromolecules for which there are no certified reference materials or methods, results can be extremely variable.[26]

Because of the heterogeneity in the way, some laboratory test results are reported in the EHR, it is important to capture sufficient metadata to enable filtering or harmonization of results when analyzing laboratory test results in EHR data. The most common standard for mapping laboratory test results is the Logical Observation, Identifiers, Names, and Codes (LOINC) which includes 93,600 (version 2.68) terms.[27] Mapping laboratory results to LOINC greatly facilitates data harmonization but LOINC does not actually include sufficient detail to resolve poorly standardized tests because it does not denote particular assays or instruments. Thus, it is critical to gather information such as performing laboratory, reporting units, reference range, and testing date, and to explore differences in results across these variables. In addition, these same principles of surveying variability and ultimately considering filtering or harmonizing data applies to every domain of clinical data.

Standards

Standards and the Electronic Health Record (Peter Gabriel)

Successful research with EHR data requires that systems be able to exchange information in a way that preserves the meaning of the data. Since different EHR systems (and even different implementations of the same system) encode data in unique ways, mapping data to an external standard is often necessary. This enables “semantic interoperability,” defined as the ability for computer systems to exchange data with unambiguous, shared meaning. While a human physician can easily understand that “pneumonia” and “lung infection” are the same, a computer cannot. Achieving semantic interoperability is the main driver for meaningful health care data exchange and aggregation at all levels (within and between organizations, regional or research networks, and even countries). It requires standardized documentation of medical knowledge and the health care process.[28]

One way to achieve this has been the development of standardized vocabularies for medical terms which often codify specific types of EHR data. For example, the 10 revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10), described in detail in the next section, consists of diagnostic and procedural codes.[29] The LOINC is an international standard for codifying health measurements, observations

(such as laboratory and imaging tests), and documents, whereas RxNorm is the standard medications terminology in the United States.[30] The Systematic Nomenclature of Medicine Clinical Terms (SNOMED CT) is the most comprehensive clinical terminology in the world and is available to the U.S. users at no cost through the National Library of Medicine's UMLS Metathesaurus. SNOMED CT is organized into concepts, descriptions, and relationships. Concepts represent a unique clinical entity or process, and have a unique identifier associated with them. Descriptions of the concepts that allow for specification of synonyms, help enable semantic interoperability through mapping of equivalent terms to a single underlying concept. Finally, the relationships describe how concepts are associated with one another.

In addition to the information content standards mentioned above, information exchange standards help to ensure that data are transmitted between systems with a consistent structure and organization. The Health Level Seven International (HL7) is an American National Standards Institute–accredited standards developing organization responsible for several widely used information exchange standards, including HL7 Messaging Standard Version 2 (HL7 V2), HL7 Clinical Document Architecture (HL7 CDA), and HL7 Fast Healthcare Interoperability Resources (HL7 FHIR, https://www.hl7.org). Similarly, the Clinical Data Interchange Standards Consortium (CDISC) develops and maintains multiple information content and exchange standards specifically dedicated to supporting clinical research activities. Finally, several common data models (CDMs) have gained widespread adoption in recent years. These specify a consistent structure for storing data in databases that can facilitate combining datasets together, as well as distributing queries across a federated network of databases, and aggregating the results. The Observational Medical Outcomes Partnership (OMOP) CDM is one such model, used worldwide for observational research.[31]

Billing Codes to Record Diagnoses (Mary R. Boland)

A primary function of the EHR is the documentation of the events that occur during routine clinical care. Therefore, EHRs contain rich information detailing the clinical encounter and the patient's experience within the health care system. Researchers often seek to repurpose these rich data sources for use in studies to understand disease burden, treatment efficacy, and health outcomes. A key feature of EHR data is the use of billing codes to record diagnoses, procedures, medications, and other billable events that occur as the patient navigates through the health care system. These codes are recorded by physicians through their encounters with EHR platforms and systems. However, in one common EHR system, EPIC (Epic Systems Corporation, Inc.), physicians do not often directly record codes, but concepts (e.g., chief complaint and diabetes) that are translated to codes later). Medical coders, along with specialized automated coders, apply hospital billing codes called the ICD codes to each patient's encounter to encode diagnoses, diseases, conditions, comorbidities, and other aspects of the patient clinical status recorded during their care. The World Health Organization releases versions of ICD billing codes that incorporate periodic updates. The versions used most frequently in the United States are v.9 (up until c.2015) and v.10 (2015-present), and a more recent version ICD-11 which has not gained traction in the United States yet but will likely be used in the next decade or so.[32]

Billing codes, such as ICD-10, are very comprehensive and include many very detailed observations that clinicians make regarding the patient during the clinical encounter. For example, there is a code for being hurt in the library (ICD-10 Y92.241—"Library as the place of occurrence of the external cause"), and there are extremely detailed codes pertaining to sporting accidents, including being burned by water skis on fire (ICD-10 V91.07XA—"Burn due to water skis on fire, initial encounter"). These codes are often used in combination with other codes (e.g., injury codes such as broken arm) to provide rich and detailed context around why the patient may have broken their arm. ICD billing codes are organized within a hierarchy comprised of a range of health topics and subtopics based on mechanisms (e.g., infectious and parasitic diseases), affected organ systems, (e.g., diseases of the respiratory system), or disease related (e.g., neoplasms) among other topics. However, the ICD taxonomy does not uniformly represent clinical reality or the interpretation of that reality by clinicians and coders.[33]

Researchers will leverage these ICD codes to define clinical phenotypes and health outcomes to study a group of patients both with (cases) and without these characteristics (controls). However, researchers should be aware that codes are often used to justify a diagnostic procedure that otherwise might not be covered by a payor. Frequently, these codes are not removed from the EHR when the diagnostic test is subsequently reported as negative. This creates the possibility of false positives in a query of EHR data for the purposes of cohort identification. In addition, some codes are not to be used in billing (such as accident detail codes) and those codes are annotated as nonbillable codes in the record systems. These "nonbillable codes" are often annotated by clinicians to provide clinical context to the billable codes and for reporting purposes after the fact. However, it is important that the majority of ICD codes are recorded and captured for the purpose of billing, and therefore, their ability to correctly define a patient's clinical phenotype or health outcome for research can be highly variable.[34] Furthermore, in some instances insurance claims are rejected due to the selection of billing codes. This can lead to altering the initial billing codes to a more refined set of billing codes (e.g., a more specific disease code vs. a more general one) that may satisfy the insurance company, but may not be as useful for research. In addition to ICD codes, there are also procedural codes annotated by the Current Procedural Terminology (CPT) codes and medication codes annotated by RxNORM, along with a specialized coding terminology for oncology called the ICD for Oncology version 3, or ICD-O-3, that captures cancer morphology, histology, and behaviors. These clinical domain-specific lists can provide additional granularity beyond traditional ICD code lists useful for answering clinical research questions. These billing data can be further enhanced through coupling with additional clinical information (signs, symptoms, radiological findings, and laboratory/result values) housed in EHRs derived from structured and unstructured clinical data (natural language processing [NLP] and image processing) to further improve specificity and sensitivity of EHR phenotyping algorithms.[35]

Ontologies (Charles Kahn)

Ontologies encode biomedical knowledge that in turn can empower mining of EHR data. An ontology specifies a domain's concepts along with the relationships among those concepts and can be used to standardize the terms used in a field of discourse. An ontology is

more than a controlled vocabulary. The connections between terms make ontologies more powerful than simple vocabularies: those connections define the concepts' semantics, or "meaning." Through those relationships, ontologies express knowledge in a form that is both computable and human readable.[36] Unlike individual terminologies and coding systems, ontologies are supported by the World Wide Web Consortium (w3c.org) standards that enable linked data sharing. Ontologies can specify part-whole relations (e.g., lingula is part of left upper lobe, which is part of left lung), subclass–superclass ("is-a") relations (e.g., adenocarcinoma is a type of carcinoma), and various other kinds of relationships. These relationships make ontology-enabled applications more powerful.[36] For example, a search for patients with "lung cancer" will identify patients with adenocarcinoma (a subtype of cancer) in the lingula (a part of a lung); the ability to reason abstractly makes the search more effective.

Biomedical ontologies—including those built on the principles of the Open Biological and Biomedical Ontology (OBO) Foundry[37]—enable sophisticated data mining from EHR data. The Systematized Nomenclature of Medicine (SNOMED) and the Disease Ontology provide a standardized representation of diagnoses. The Drug Ontology relates medications to their active ingredients; the Chemical Entities of Biological Interest links those ingredients to their biological roles. Thus, a medication record associated with a health care encounter can be used to infer diagnosis and intervention. The results of laboratory tests can be transformed into statements about a patient's phenotype.[38] Through the encoded relationships of terms representing the data, ontologies provide explicit context and semantic complexity that allows one to achieve a more accurate understanding of EHR data and to use that data more effectively.[39,40]

Common Data Models (Robert Grundmeier)

Among the many challenges that arise with EHR data are that the structures are often vendor specific, may differ across health systems even when the same vendor's product is in use, and are not organized in ways that are useful for typical research analyses.[41] Each health system typically employs highly skilled data analysts who work diligently to transform the EHR's complex data structures into formats that are more usable for research. Without a Common Data Model (CDM), these efforts typically must be repeated for each project and within each health system that may be participating in a multisite research endeavor. A more efficient approach is to consistently use a dataset with a standardized data structure across many projects. These standardized data structures are typically referred to as a CDM. Although the structures of CDM's are necessarily more complex than a data model constructed for a specific research question, they offer some significant advantages as follows:

- They are much simpler than the raw data structures of the EHR.
- Users who move from one organization to another can very quickly begin working with a CDM that they are familiar with since the data structures are the same.
- There are often communities working collaboratively to solve recurring analytic challenges together.

- Communities often provide tutorials, code examples, and forums to propose improvements to the data model or obtain extra help when needed.
- Time and money are saved by eliminating nonvalue-added work to prepare, harmonize, and learn a project-specific data model.42

There are, however, a few disadvantages to CDM's as mentioned below:

- There isn't just one, so evaluation and selection of a model needs to occur which is a project unto itself.
- A CDM requires significant effort to implement, particularly mapping and validating your EHR data.
- A CDM requires ongoing maintenance to ensure that it stays current with changes to the CDM, as well as the underlying EHR data, feeding the model.
- They do not accommodate all data in the EHR, thus specific needs may require extensions to the CDM.

A review of several CDMs identified the OMOP CDM as particularly flexible for addressing a variety of research questions involving EHR data.[43] The OMOP data model is maintained by the Observational Health Data Sciences and Informatics (OHDSI) community which includes a large and active community of collaborators.[31] Researchers interested in patient-centered outcomes research may find the PCOR-Net CDM useful in part through research opportunities facilitated by the Patient-Centered Outcomes Research Institute (PCORI).[44] For researchers seeking to combine biologic data (e.g., gene sequence or expression data) with EHR data, the i2b2 data model and related research tools may be most helpful.[45] Before developing a custom-made data model, researchers should always consider whether an existing CDM can adequately organize their research datasets. In addition, it is important to note that existing CDMs contain metadata from medical or administrative sources and have not included metadata data from other disciplines such as nursing, physical therapy, social work, or case management, nor do they include the consumer perspective. This gap should be addressed as current CDMs are updated and new ones proposed. Increasingly, there is also an expectation among funding agencies that researchers use a CDM to ensure the research datasets can be made widely available at the conclusion of the funding period to maximize the value of publicly funded research efforts.

Governance

Data Governance and the Electronic Health Record (William Hanson)

While the term data governance has been familiar in other industries for years, its relevance to medical data has only become apparent recently, with the broad adoption of electronic medical records, which generate a mix of structured and unstructured data. Medical data historically comes in a variety of "dialects," varying from institution to institution, vendor to vendor, and even from instance to instance of the same vendor's implementation. Data from the electronic record may also be concatenated with other patient datasets into complex data assemblages consisting of administrative, demographic, physiologic, laboratory, descriptive text and images, and the governance of these datasets requires input from a range of

disciplines and domains.[46] The analysis of such data for research may require expertise in NLP, data normalization, and other specific skills. A national experiment wherein data pertaining to clinical care, practice organization, and patient experience were combined to assess the quality of care provided by individual family practitioners was undertaken in the United Kingdom's National Health Service in 2004 (reference in comment). Financial rewards were assigned to high quality care amounting to as much as $77,000 dollars ($42 pounds at the then-exchange rate) and the underlying data were derived from the practitioner's computers and patient surveys. This program both relied upon and illustrated the need for increasing digitization of medical data, as well as attention to its accuracy. The interdependency of medical datasets also increasingly requires coordinated stewardship and governance, so that changes in one system are made with visibility to others and with appropriate adjustments in linked systems: changing the name or structure (i.e., the adoption of a new ICD release) of an element may have profound consequences for downstream systems. And as genetic or genomic data are integrated into datasets, highly sensitive, and unalterable patient-specific data becomes manifest in a digital format and necessitates special security, and ethical considerations.[47–49] Responsible data stewardship and rigorous data governance are increasingly essential to ensure patient privacy and data integrity.

In addition to the above principles, medical data are increasingly recognized to be of tangible value, and as with other industries, the use and or misuse of patient data can have profound financial, reputational or regulatory impacts on an organization. The acquisition of troves of patient data may be the explicit or hidden agenda in the business models of new or incumbent software vendors, as has become apparent with agreements between the United Kingdom's National Health Service and Deep Mind or the University of Chicago and Google.[50–53] Effective governance and stewardship mandate a thoughtful review of the potential consequences of both data breaches, as well as transactions in which identifiable patient data are exchanged with another party.

Data Privacy (Yong Chen)

In the last two decades, there is a growing number of research networks embedded in health care systems, such as PCORnet and OHDSI, with the goal to accelerate biomedical discoveries. These research networks have provided an unprecedented opportunity for multicenter clinical research. As multicenter health care data are stored at different locations and patient-level information is often protected by privacy regulations and rules, direct data integration across multiple data centers is often not feasible or requires large amounts of operational effort. There have been great efforts on developing novel methods to allow multicenter analysis while allowing protection of patient privacy.[54] One of the most widely adopted methods is the divide-and-conquer strategy, such as meta-analysis and its multivariate extensions, which is commonly adopted in OHDSI, where investigators at distributed sites run a prespecified model and report the analysis results (such as estimated effect sizes and their variances) to be combined via meta-analysis models. Such procedure works well in many situations but can lead to nonnegligible biases in situations involving rare exposure or rare outcomes, for example, pharmacovigilance studies of rare drug adverse events. The second set of methods are iterative distributed regressions, where investigators at distributed sites run a prespecified model and communicate with a central site iteratively

to update their results.[55] This approach leads to the identical result as if the individual patient-level data were pooled together and the prespecified model was run on the pooled data. It has been deployed in research consortia such as pScanner.[56] Another set of methods is the communication-efficient distributed algorithms, seeking a trade-off between bias and communication efficiency, includes one-shot distributed regression algorithms for binary, and time to event outcomes.[57,58] This approach aims to minimize the number of communications among investigators at different clinical sites, while seeking estimates that are close to the analysis results from pooled data. Although these methods require communication of more statistics than traditional meta-analysis, the communicated statistics are summary statistics and are often privacy preserving. One common limitation of the latter two approaches is that the data are assumed to be homogeneously distributed across clinical sites. Future work is needed to extend the distributed algorithms to heterogeneous data. In addition, missing data present some unique challenges in a distributed data setting, while some privacy-preserving missing data methods have been developed, this remains a nascent area of research.[59,60] Furthermore, sharing of genetic data and biobank data are important to facilitate large-scale genetic studies, but special attention is needed to protect patient privacy and confidentiality.

Deidentification (Tessa Cook, Danielle Mowery, Nebo Mirkovic, and Laura Fluharty)

To preserve the privacy of individuals and restrict knowledge of their immediate social circles (e.g., relatives, household, and employer) while conducting research with EHR data, the data must first be deidentified. The federal Health Insurance Portability and Accountability Act (HIPAA), codified as 45 CFR §160 and 164 (US Code of Federal Regulations) and the Common Rule, prescribes two options for deidentification of clinical data: (1) removal of 18 “safe harbor” identifiers such as person’s name, address, date of birth or other unique identifying information that is considered protected health information (PHI); or (2) certification by an expert that the risk of reidentification of an individual from the data are low.[61] Most EHR research follows the first approach. For epidemiologic and population health research, a limited dataset may contain some identifiers needed to model cohorts over time (ages and dates) and geographical space (zip codes).

Obfuscation is another approach that reduces the likelihood of reidentification of any person, by transforming individual data points into categories to reduce uniqueness of data. However, none of these techniques guarantee true anonymization which eliminates any link to the individual’s identity.[62] Recent studies have demonstrated that by combining individual attributes in deidentified datasets with publicly available data, individual uniqueness can be exploited to reidentify individuals.[63] Because anonymization can decrease information content of the dataset, differential privacy methods are sometimes used in research because they divulge insightful characteristics of cohorts while not disclosing information about any particular individual. Furthermore, several tools exist for redacting PHI elements from EHR data; however, their performance depends on the types of PHI in the data and on how the data are documented.[64] Data recorded in discrete/structured form has little variability and can be more readily redacted or obfuscated. Examples include demographics (e.g., names, age, and date of birth), dates and times of clinical events (e.g., admission, discharge), and billing/administrative information (e.g., medical record number). However, unstructured

clinical text can be highly variable with regard to frequency, terms, formats, and types of PHI documented. For example, pathology results may solely contain the patient's name, and dates of clinical events; in contrast, discharge summaries often contain rich information about family members, patient's date of birth, treatment dates and locations, as well as demographics. Other data modalities present unique challenges to preserving privacy. Head and neck computed tomography (CT) scans can contain distinguishing facial features. By definition, genetic data are uniquely linked to an individual's identity, and the growing number of public repositories containing genetic information has inadvertently created new challenges to preserving participant autonomy.[65] Additional restrictions may occur for particular sensitive populations (mental health, AIDS/HIV, etc.), as governed by individual state laws. Data sharing among organizations (e.g., in the context of multicenter studies) requires additional legal controls such as data use agreements or business agreements, as well as appropriate technology for secure data exchange and storage.

Data Integration

Data Integration and the Electronic Health Record (Jeffrey Morris and Dokyoon Kim)

Correlative analyses of EHR data with other external data sources can provide significant insights, especially in identifying important explanatory factors, as well as discovery of prognostic or predictive biomarkers that may have translational value for precision therapy. There are many types of external data useful for these purposes, including geographic, socioeconomic, geospatial measurements of air quality or climate variables, genetic and multiplatform genomic data, radiological and other imaging data, and wearable device or mobile health data including mobility or activity levels, heart rate, or blood pressure data. Many of these external data are high-dimensional and complex, raising considerable informatics and analytical challenges. One important decision is whether to use a feature extraction approach, computing and analyzing simple summaries from the complex data, or try more advanced modeling approaches, such as functional data analysis, that focus on modeling the entire data structure as a complex object. Feature extraction is computationally efficient and can work well if the features extracted contain the salient information contained by the data, but can miss out on key insights if not captured by the features. Functional data modeling approaches have potential to find insights missed by feature extraction approaches, but their complexity and computational intensity can make them more difficult to implement. Another key factor to consider is multiple testing. If many external variables are screened to find associations with clinical outcomes, it is important to use training/testing strategies, permutation tests, or multiple-testing adjustment to account for the multiple testing and prevent reporting of spurious false-positive results.

Pecoraro et al also point out that the use of EHR has the advantage of managing standardized data and documents already integrated in a health infrastructure that can be easily extended comprising ad hoc information systems.[66,67] Additionally, Botsis et al emphasized that automatic or advanced data validation and flexible data presentation tools should be developed to ensure information integrity. Effective strategies (e.g., new tools, better classification systems, and others), for secondary use of EHR data, could be

also accumulated from case studies and shared with the research community as the best practices.[68]

Data Warehouses as a Challenge for Research with Electronic Health Record Data (Peter Gabriel and James Beinlich)

Health care data warehouses containing EHR data present a significant challenge for research use due to factors relating to the history of EHR's and the state of database technology and the maturing of analytics in general. EHRs are optimized for the efficient capture and storage of the billions upon billions of individual, patient-level transactions that support the highly complicated clinical and administrative processes of a modern health care system. Their origins are rooted in the automation of administrative processes like registration, scheduling, and billing. Historically, EHRs provided these central functions, while richer clinical information existed in specialty-specific ancillary systems that exchanged data with the EHR through electronic interfaces. This has been changing, however, as the larger, market-dominating EHR systems have evolved into fully integrated platforms providing deep support for a broad set of ancillary and specialty-specific areas.

The resulting complexity of these systems is striking. EHR databases typically contain hundreds of thousands of data tables and millions of individual data elements. In addition, the way the systems are configured and used (and the resulting patterns of data capture) can vary significantly over time and across different areas of a large hospital or health care system, let alone across institutions. This makes it a formidable challenge to extract, transform, and summarize the data into a warehouse environment that correctly preserves the meaning of the original data. And even with careful design, proper interpretation of warehouse data can be difficult for secondary users who lack a detailed working knowledge of the operational workflows that led to its original capture. Data in the EHR are also limited by the data that are entered into the EHR. As Weiskopf et al pointed out, the completeness of EHR data for secondary uses is lower than one might expect for a variety of reasons that fall into four general categories: (1) documentation, (2) breadth, (3) density, and (4) prediction.[69] In an effort to address completeness issues that affect secondary use, organizations need to balance increased need for data with the pushback from busy clinicians that already feel overwhelmed by data entry requirements into the EHR. As Kroth et al point out in a survey of 282 clinicians, 86.9% cited excessive data entry requirements as one of the most prevalent concerns about EHR design and use.[70]

It is also important to point out that until recent changes in the economic health care model (i.e., accountable care organizations and value-based care), many data points were only within the capturing organization's immediate scope of care (i.e., a particular hospital stay or clinic visit), leaving gaps in the complete picture of a patient's care. As health care providers are becoming more accountable financially for patient outcomes and utilization beyond their direct scope of care, more external data points are being captured in each institution's clinical systems.

From a technological perspective, data warehouses have historically been designed to overcome the limitations (such as complexity and performance) of operational databases which are optimized to support transactional volume processing. Much like the history of

EHR's, reporting databases and systems have traditionally supported administrative and operational reporting activities (capacity, billing, etc.) as opposed to population health and clinical needs which has resulted in functionality gaps to address things to support accountable care organizations and value-based care. Until recently, data warehouses have been massive, relational databases that do a great job of harmonizing vast amounts of data from a variety of operational stores, but they require significant up-front development and implementation to optimize performance. This approach forces organizations to assume in advance what questions will be asked of the warehouse to design and implement it. Over time, organizations have learned that this, while it provides significant value, is also very inflexible (some changes require rearchitecting of the warehouse) and results in long timelines to introduce enhancements. As the move to cloud technology, in-memory processing, and flexible data architecture have evolved, the challenges of legacy data warehouses are quickly being addressed. This is critical to address the fast-paced and everchanging nature of data analytics and biomedical informatics.

Clinical Care

Patient Surveys for Precision Medicine (Marylyn Ritchie)

The promise of precision medicine relies on our ability to integrate all relevant risk factors for disease which include clinical risk factors that can be captured in an EHR, but also environmental exposures, behavior, and social determinants of health (SDOH). Extracting the relevant clinical risk factors from an EHR is done on a regular basis. However, a significant amount of data that are needed for precision medicine are simply not captured thoroughly in an EHR. A 2015 report from the National Academy of Medicine (NAM) identified multiple domains and measures of SDOH that are important for health and disease that should be better captured in an EHR but are not captured by clinical providers.[71] These domains include: sociodemographic factors (education, employment, and financial resource strain), psychological factors (health literacy, stress, negative mood, and affect), behavioral domains (dietary patterns, physical activity, tobacco, and alcohol use), social relationships (social connections and/or isolation and exposure to violence), and neighborhoods and community characteristics.[71] While advances have been made since the time this report was written in 2015, EHRs still lack most of these measures. This gap could be addressed by including the nursing, physical and occupational therapy, and social work communities in the endeavor to enrich the EHR with these data. Clinical researchers focus on SDOH factors in research projects; the consequence is that these data are relegated to sections of the EHR that are not routinely accessed by physicians or researchers or that they reside in research databases outside of the EHR.

In some precision medicine research programs like the United Kingdom Biobank, which is a population-based biobank, data on physical activity, dietary information, and social/behavioral measures are abundant. Much of these data are collected via participant-reported surveys.[72] It seems that a powerful approach for precision medicine research in the future would include routinely conducting patient-reported surveys to supplement the EHR. If informatics tools, algorithms, and workflows to enable the addition of robust patient-participant reported data to capture environmental exposures, behavior and SDOH

data can be leveraged to supplement EHR data, these additional data will greatly enhance opportunities for precision medicine research using EHR data

One way to conduct these surveys in patients from health systems to link with the EHR data for research is through the use of the PhenX Toolkit (https://www.phenxtoolkit.org/).[73] The PhenX Toolkit is an online catalog of standard measurement protocols. Currently, there are over 800 measurement protocols in the Toolkit addressing 25 research domains. PhenX protocols cover a broad scope of research domains (e.g., demographics, cardiovascular, diet and nutrition), while collections provide depth in specific areas (e.g., substance abuse and addiction research and mental health research). For example, PhenX surveys were used in conjunction with EHR data at the Marshfield Clinic. Surveys were mailed to participants through the Marshfield Personalized Medicine Research Project (PMRP).[74] A total of 36 measures from the PhenX Toolkit were chosen within the following domains: demographics, anthropometrics, alcohol, tobacco and other substances, cardiovascular, environmental exposures, cancer, psychiatric, neurology, and physical activity, and physical fitness. The results of this study and the high response rate highlight the utility of the PhenX Toolkit as a path forward to collect valid phenotypic data that can be used to augment the data available in the EHR.

While it is clear that the integration of SDOH and environmental data with the EHR would improve health care and research, it is important to note that the lack of behavioral and SDOH data is pervasive at an international level, and not unique to health systems and EHRs. For example, an analysis of the main indicators to assess the quality of child care in 30 European Union (EU)/(European Economic Area) EEA countries highlights that the focus continues to emphasize clinical indicators and does not include both individual and community wellbeing factors (Luzi et al).[75]

Clinical Decision Support Systems (Kathryn Bowles and Michael Draugelis)

Building, validating, and implementing clinical decision support (CDS) systems from EHR and administrative data holds many challenges, especially data availability and the quality of that data. Then, once implemented, CDS system maintenance brings another set of unique challenges because health system data are noisy and dynamic. Human-entered data contains temporal noise from corrected or deleted data over time, creating a changing state of knowledge that health system EHR does not often capture. The missing temporal information makes it difficult for researchers or engineers to reproduce or predict system performance. This hidden technical debt results in a brittle CDS system whose accuracy is continually degrading. Data governance can help by limiting who, when, and how data are accessed, as well as reducing incomplete or inaccurate documentation,[76,77] and variation in data fields.[41] In addition to data governance, the CDS maintenance challenges call for robust active data monitoring, root-cause analysis, and routine system updates.[78,79] Having the skilled workforce to manage data quality, extraction, and CDS maintenance is critical and may alone be difficult to achieve.

Clinical Research

Registries (Tessa Cook)

Patient registries are a powerful mechanism for collecting EHR data for quality improvement and research. They can be constructed within a single hospital or health system, or data collected from the EHRs of multiple hospitals and health systems to facilitate large-scale, population-based research. Within a single health system, constructing a registry requires careful analysis of the data elements to be used to identify patients who belong in the registry. Across multiple contributing systems, setting up a registry is a complex task that requires coordination between experts in information security, patient privacy, data science, and statistics, in addition to research and health care. One of the biggest challenges in collecting data from different EHRs and harmonizing it for addition to a registry is the fact that these different EHRs use unique labels and organizational schemes for their data.[80] No preexisting harmonization mechanism exists. As a result, every new registry that is built requires the same steps to be performed again. To address this particular challenge, some have explored empowering patients to contribute their data directly to registries.[73,81] However, this approach has its own challenges that are beyond the scope of this discussion.

Registries may collect data for the same patients in a phased approach to make the process of contributing to the registry appealing to the participating sites. However, this results in another major challenge, namely, establishing and maintaining the accuracy of the data within the registry. Because of concerns surrounding patient privacy and data security, actual patient identifiers are almost never contributed to a registry. Instead, a mapping table between the patient's true identity and an anonymized identifier is often maintained at the contributing site. It can be onerous for a site to access a specific medical record for periodic updates to the registry.[82] Additionally, errors in the mapping can lead a site to reference the wrong patient's record when updating the registry, thereby decreasing the accuracy of the data within the registry.

Study Design (Blanca Himes and Rebecca Hubbard)

Epidemiologic studies and clinical trials are carefully designed in advance of data collection to address specific hypotheses. In contrast, studies that use EHR data for secondary purposes rely on data generated by a complex process involving patients choosing to interact with a health care system, and health care providers and administrators documenting these interactions for the purposes of recording clinical care and billing.[83] Although EHR data can be made to superficially resemble a traditional cross-sectional or longitudinal study, the underlying data generation process is starkly different, and this must be considered when performing research. Novel and actionable health insights can result from carefully designed EHR studies, but if they lack domain knowledge, naïve researchers may design a study that only yields findings that are tautological or reflect confounding (e.g., smoking history is associated with chronic obstructive pulmonary disease (COPD) and hemoglobin A1C tests are associated with diabetes). Health care encounters are sporadic and may not capture highly relevant health information that occurs between visits, including encounters with providers in other health systems. Additionally, a wide range of information on

behavioral, social, economic, and environmental factors that influence health are often not recorded in EHRs.[84,85] Sicker individuals tend to be overrepresented in EHR-derived datasets because they more frequently visit health systems to manage multiple comorbid conditions and for urgent care resulting from these conditions.[86,87] Even data recorded during a single encounter imperfectly represents health processes of interest to researchers due to uncertainty of tests used for health assessment and errors in phenotypes assigned based on what is available in the EHR. For example, phenotypes can be erroneous due to failure to record research-relevant information that was not necessary at the point of care, or coarse assignment of conditions based on billing codes, rather than results of a full clinical assessment. Moreover, the timing of phenotype detection may lag behind the health events themselves, with this lag varying across patients depending on the timing and type of health care encounters the patient experiences. As a result, EHR-derived phenotypes are an imperfect reflection of patient health, and methods for analysis of EHR data must appropriately account for the complex process that connects them to the underlying truth.

Statistical Issues as a Challenge for Research with Electronic Health Record Data (Marylyn Ritchie and Qi Long)

While EHRs contain an enormous amount of rich, longitudinal data which can be very powerful for observational research studies, there are several statistical issues that need to be considered and overcome to generate robust, unbiased findings from analysis of EHRs data. First, EHR data are fraught with measurement errors and most notably phenotyping errors. It has been widely acknowledged that EHR may not contain sufficient data for some clinical endpoints and particularly, ICD codes, developed for billing purposes, that do not always accurately capture all medical conditions. For example, it is challenging to define progression-free survival in oncology using EHR data. While advanced statistical and ML methods have been developed to improve phenotyping accuracy using EHR data,[88] the resulting “predicted” phenotypes still have some uncertainty that needs to be adequately accounted for in subsequent statistical analysis.[89] Another statistical issue in EHR data is that of variation in the time intervals of measurements and diagnoses across patients. Unlike clinical trials or prospective epidemiological cohorts with regularly scheduled follow-up visits, because the patient data in an EHR is populated when they visit the health system, there is irregularity in the time intervals for different patients. The frequency of patient visits and lengths of intervals are often associated with patient’s underlying health condition. For some patients, they have routinely scheduled appointments, while for others, there are only sporadic and occasional visits. This variability can lead to challenges when trying to perform longitudinal and/or time series analyses and can lead to biased results if not adequately accounted for in these analyses. There are ongoing efforts on addressing these and other challenging statistical issues associated with EHRs data.[90]

Missing Data (Blanca Hubbard)

Inherent in clinical care is a bias toward ordering tests as needed for better diagnosis, and performing procedures and ordering drugs in accordance with patient needs. Additional factors, such as type of insurance, can influence what is ordered for patient care and eventually recorded in the EHR. Because prespecified protocols do not guide which EHR data elements should be collected, unavailability of tests or other information in EHRs does

not imply “missing” in the traditional sense of failure to collect data despite an intention to do so. Rather, the timing and type of measures available in the EHR are driven by clinical and administrative—not research—needs, resulting in inconsistent assessment of patient characteristics. Even when rigorous effort is exerted to extract information from EHRs, including via manual chart abstraction, text mining, and NLP, some data elements will be unavailable for some patients simply because not all measures of interest to researchers were recorded. Similarly, presence of data in the EHR can reflect purposeful decision-making, as demonstrated by a retrospective study that found that laboratory orders were better predictors of outcomes than actual test results.[91] Because researchers typically have no control over missing or inconsistently collected EHR data, it is critical to understand the causes of missingness and adopt methodology that appropriately accounts for the missingness mechanism, realizing that in some cases, a problem of interest may simply not be addressable.

Many methods have been developed for the analysis of missing data which occurs in all studies, regardless of their design. Each approach relies on assumptions about the underlying missing data mechanism, the most common of which is that data are missing at random (MAR). In the case of MAR, the probability that a measure is missing can be predicted on the basis of observed variables. Popular methods such as multiple imputation and inverse probability weighting make the MAR assumption. Because EHR data are often not MAR, these popular methods may not be appropriate to deal with missing EHR data. In the case of missing not at random (MNAR), the probability that a measure is missing is associated with the unobserved value itself or other unobserved patient characteristics. For EHR data, MNAR results when individuals who are sicker have more encounters, and thus, more measurements and types of measures recorded. For cognitively impaired patients, missing data may be meaningful and related to the condition because the patient was unable to provide the information. Haneuse and Daniels[92] suggested that consideration of why EHR data were observed and recorded can guide researchers to appropriately use missing data methods. Modeling the mechanism underlying the observation process based on an understanding of data provenance is more likely to lead to correct specification and hence, valid analysis, than the traditional approach to missing data which focuses on specifying reasons that data are missing.

Machine Learning and the Electronic Health Record (Ryan Urbanowicz and Michael Draugelis)

Machine Learning (ML) holds great promise as a tool for the detection of patterns and associations, particularly when the number of potentially predictive variables become large, and there is a need to consider complex multivariate relationships. However, ML is far from a silver bullet, and its effective use relies on many factors, for example, data dimensions (number of variables and instances), signal-to-noise ratio, selecting appropriate methods, optimizing hyperparameter settings, and eliminating or accounting for bias. EHR data include extensive yet noisy clinical notes, diagnosis and procedure codes, laboratory test measurements and results, medication prescriptions, as well as biomedical images such as computerized tomography (CT), magnetic resonance imaging (MRI), or pathology images where relevant. Thus, there are some specific considerations when applying ML to EHR

data due to the biases inherent in EHR. First, applying ML to EHR always constitutes a secondary data analysis, meaning that ML is being applied to answer questions that the data were not collected to specifically answer.[93] This ultimately increases the amount of noise and the diversity of bias sources that can negatively impact ML performance. Second, EHR data contain longitudinal health information for an individual patient at the point of care. Thus, when we construct a three-dimensional matrix that consists of patients, a set of variables, and time, the matrix is extremely sparse. This nature of sparseness of EHR data makes ML very difficult to train robust models for the outcome prediction. Third, disease phenotypes are often not definitively defined or validated in EHR which has led to significant interest in computational phenotyping, where phenotypes are themselves derived through the application of ML methods.[94] Fourth, EHR data typically offer a mixture of variable types (e.g., binary, ordinal, quantitative, and categorical). Some ML algorithms can favor certain variable types when mixed together in a dataset.[95] This is one of many potential sources of bias that can impact ML performance.[19] Lastly, ML predictive modeling requires structured data to train upon. EHR data often include unstructured text or other variables that are poorly suited to ML algorithm modeling. Thus, prior to ML modeling, EHR data often require preprocessing via NLP and/or feature engineering. Despite these considerations, we believe that ML will have a positive impact on many aspects of biomedical research with EHR.

Overlap of the Subfacets

The matrix shown in Table 1 illustrates the overlap between the various subfacets as presented and discussed in the previous sections, and indicated here by the section number. There are several facets that are remarkable for the number of intersections between subfacets. These include usability, data quality, data standards, and data integration, and reflect the importance of these facets in understanding the use of the EHR for clinical care and research (Table 1).

Conclusion

The value of EHRs was clearly demonstrated as health systems around the world relied on patient data to understand and respond to the coronavirus pandemic and resulting COVID-19. Some of the lessons learned about our health information technology infrastructure and informatics methodologies for dealing with this infectious disease emergency have been recently reviewed.[96] One of these challenges relates to the integration of data across health systems to achieve better statistical power and to identify clinical patterns and associations which are more likely to generalize across different patient populations. This integration can happen in a centralized manner as demonstrated by the National COVID Cohort Collaborative (N3C) consortium[97] or in a federated manner as demonstrated by the Consortium for the Clinical Characterization of COVID-19 by EHR(4CE)[98] or the OHDSI consortium.[99] These data sharing efforts and others have revealed that, while we have made a lot of progress on CDMs and standards for data integration, there is much work left to do.

The goal of this review was to summarize the challenges of deriving clinical value and research insights from EHR data. We covered the data harmonization, integration, and storage challenges faced by the COVID-19 research consortia, as well as many other challenges, such as inflexibility of EHRs, data privacy and security, fairness and bias, data quality and variability, and analytics including study design, statistics, and machine learning (ML). While each of these topics merits intense study, it is important to step back and think about how they relate to one another in the context of a learning health system where basic science and clinical research are closely integrated with biomedical informatics and clinical care to improve the health of individuals and the health care delivery process.[100] Central to the learning health system is the availability of EHR data in an integrated, standardized, fair, and user friendly manner. Basic scientists and clinical researchers need to be able rapidly query and get access to clinical data in a self-service manner. CDMs and ontologies can play an important role in this self-service model by allowing for natural language queries of the data for cohort discovery. Such queries enable anyone to query the data thus bypassing the need to engage a data wrangler who are often in short supply with a health system. Self-service identification of cohorts and the seamless transfer of clinical data to a secure server could greatly reduce the time this normally takes from weeks or months to hours or days. This single time reduction could greatly accelerate the generation of research results and new knowledge which could benefit a learning health system. Thus, we recommend learning health systems invest in self-service data access as a way to accelerate the discovery process.

An additional slow step in the learning health system process is data cleaning and quality control. As we have discussed, clinical data from EHRs are complex and noisy. Assessing and dealing with bias, missingness, extreme values, etc., can consume a lot of time with large datasets with hundreds or thousands of variables. It is not unusual for the data cleaning phase of the analysis to consume several months or longer. When combined with inefficient data access, it is not unusual for 12 months to pass before statistical and ML analyses can begin to address a clinical question. Unfortunately, there has not been much progress on automating the data cleaning and quality control process for EHR data. Reducing the time it takes to clean data from several month to several days or weeks would eliminate a major bottleneck. As such, we recommend federal, private sector, and health system investment in automated data cleaning methods.

A major barrier on the data science front is access to statisticians and informaticians to collaborate on statistical and computational analyses. This is particularly true given the rise in demand for artificial intelligence (AI) and ML expertise to identify patterns in big clinical data. Unfortunately, advanced statistical methods and many AI and ML methods are beyond the reach of nonexperts. This is due to their mathematical and algorithmic complexity, computational demands, implementation decisions, such as parameter setting, and the lack of user friendly software. A promising development in this space is automated ML (AutoML) which is designed to take much of the guesswork out of implementing ML.[101] AutoML has the potential to bring powerful ML methods to nonexperts. The key will be to move AutoML software from the command line to user friendly graphical interfaces. An additional goal is to automate the interpretation of models generated by AutoML. There is much work to be done with AuoML. However, it has the potential to greatly accelerate

discovery by putting powerful computational methods in the hands of anyone who want to do so. We recommend federal, private sector, and health system investment in AutoML methods and their interpretation and explainability.

What COVID-19 has revealed to us is the urgency of making progress on these challenges, so that access and use of EHR data can become routine and efficient. We have made several recommendations related to accelerating scientific discovery by developing and deploying informatics methods and tools for self-service data access and automated data cleaning and ML. These steps alone could save users months of time for each study. It is important to not lose sight of the many other pieces of the learning health system puzzle. For example, advancing a ML model into the clinic through clinical decision support can take years for model validation and eventual adoption as part of a clinical workflow with decision support. Informatics has a very important role to play in this part of the learning health system process as well. It is our hope that this review will motivate researchers with quantitative and computational skills to tackle these challenges and will motivate new policies and funding to make this a priority.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number: UL1TR001878. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Office of the National Coordinator for Health Information Technology. Office-based physician electronic health record adoption. Accessed December 8, 2020 at: dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php
2. Office of the National Coordinator for Health Information Technology. Percent of hospitals, by type, that possess certified health IT. Accessed December 8, 2020 at: dashboard.healthit.gov/quickstats/pages/certified-electronic-health-record-technology-in-hospitals.php
3. Artis KA, Bordley J, Mohan V, Gold JA. Data omission by physician trainees on ICU rounds. Crit Care Med 2019;47(03):403–409 [PubMed: 30585789]
4. Shenvi EC, Feupe SF, Yang H, El-Kareh R. “Closing the loop”: a mixed-methods study about resident learning from outcome feedback after patient handoffs. Diagnosis (Berl) 2018;5(04):235–242 [PubMed: 30240357]
5. Khairat S, Coleman C, Newlin T, et al. A mixed-methods evaluation framework for electronic health records usability studies. J Biomed Inform 2019:94:103175 [PubMed: 30981897]
6. Cohen DJ, Dorr DA, Knierim K, et al. Primary care practices’ abilities and challenges in using electronic health record data for quality improvement. Health Aff (Millwood) 2018;37(04):635–643 [PubMed: 29608365]
7. Bristol AA, Nibbelink CW, Gephart SM, Carrington JM. Nurses’ use of positive deviance when encountering electronic health records-related unintended consequences. Nurs Adm Q 2018;42(01):E1–E11
8. Gephart S, Carrington JM, Finley B. A systematic review of nurses’ experiences with unintended consequences when using the electronic health record. Nurs Adm Q 2015;39(04):345–356 [PubMed: 26340247]

9. Friedman A, Crosson JC, Howard J, et al. A typology of electronic health record workarounds in small-to-medium size primary care practices. J Am Med Inform Assoc 2014;21(el):e78–e83 [PubMed: 23904322]
10. Schiff GD, Zucker L. Medical scribes: salvation for primary care or workaround for poor EMR usability? J Gen Intern Med 2016;31(09):979–981 [PubMed: 27412424]
11. Flanagan ME, Saleem JJ, Millitello LG, Russ AL, Doebbeling BN. Paper- and computer-based workarounds to electronic health record use at three benchmark institutions. J Am Med Inform Assoc 2013;20(e1):e59–e66 [PubMed: 23492593]
12. Hysong SJ, Sawhney MK, Wilson L, et al. Provider management strategies of abnormal test result alerts: a cognitive task analysis. J Am Med Inform Assoc 2010;17(01):71–77 [PubMed: 20064805]
13. Menon S, Murphy DR, Singh H, Meyer AND, Sittig DF. Workarounds and test results follow-up in electronic health record-based primary care. Appl Clin Inform 2016;7(02):543–559 [PubMed: 27437060]
14. Zahabi M, Kaber DB, Swangnetr M. Usability and safety in electronic medical records interface design: a review of recent literature and guideline formulation. Hum Factors 2015;57(05):805–834 [PubMed: 25850118]
15. Roman LC, Ancker JS, Johnson SB, Senathirajah Y. Navigation in the electronic health record: a review of the safety and usability literature. J Biomed Inform 2017;67:69–79 [PubMed: 28088527]
16. Reid PP, Compton WD, Grossman JH, Fanjiang Geds. National Academy of Engineering (US) and Institute of Medicine (US) Committee on Engineering and the Health Care System. Building a Better Delivery System: A New Engineering/Health Care Partnership Washington, DC: National Academies Press (US); 2005
17. Health information technology: standards, implementation specifications, and certification criteria for electronic health record technology. 2014 edition; revisions to the permanent certification program for health information technology. Accessed June 17, 2021 at: https://federalregister.gov/a/2012-20982
18. Ratwani RM, Fairbanks RJ, Hettinger AZ, Benda NC. Electronic health record usability: analysis of the user-centered design processes of eleven electronic health record vendors. J Am Med Inform Assoc 2015;22(06):1179–1182 [PubMed: 26049532]
19. 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):1544–1547 [PubMed: 30128552]
20. Kearns M, Neel S, Roth A, Wu ZS. An Empirical Study of Rich Subgroup Fairness for Machine Learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19. Association for Computing Machinery; 2019100–109 Doi: 10.1145/3287560.3287592
21. Gijsberts CM, Groenewegen KA, Hoefer IE, et al. Race/ethnic differences in the associations of the framingham risk factors with carotid IMT and cardiovascular events. PLoS One 2015;10(07):e0132321 [PubMed: 26134404]
22. Eneanya ND, Yang W, Reese PP. Reconsidering the consequences of using race to estimate kidney function. JAMA 2019;322(02):113–114 [PubMed: 31169890]
23. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169(12):866–872 [PubMed: 30508424]
24. Ehrmeyer SS, Laessig RH. Has compliance with CLIA requirements really improved quality in US clinical laboratories? Clin Chim Acta 2004;346(01):37–43 [PubMed: 15234634]
25. Greg Miller W, Myers GL, Lou Gantzer M, et al. Roadmap for harmonization of clinical laboratory measurement procedures. Clin Chem 2011;57(08):1108–1117 [PubMed: 21677092]
26. Christenson RH, Jacobs E, Uettwiller-Geiger D, et al. Comparison of 13 commercially available cardiac troponin assays in a multicenter North American study. J Appl Lab Med 2017;1(05):544–561 [PubMed: 33379796]
27. Huff SM, Rocha RA, McDonald CJ, et al. Development of the logical observation identifier names and codes (LOINC) vocabulary. J Am Med Inform Assoc 1998;5(03):276–292 [PubMed: 9609498]
28. Cornet R, Chute CG. Health concept and knowledge management: twenty-five years of evolution. Yearb Med Inform 2016;1(9312666):(Suppl 1):S32–S41 [PubMed: 27488404]

29. Centers for Disease Control and Prevention. International Classification of Diseases, 10th Revision (ICD-10). Accessed June 17, 2021 at: https://www.cdc.gov/nchs/icd/icd10.htm
30. Bodenreider O, Cornet R, Vreeman DJ. Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. Yearb Med Inform 2018;27(01):129–139 [PubMed: 30157516]
31. Hripcsak G, Duke JD, Shah NH, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574–578 [PubMed: 26262116]
32. World Health Organization. International Statistical Classification of Diseases and Related Health Problems (ICD). ICD-11 Accessed June 17, 2021 at: https://www.who.int/classifica-tions/icd/en/
33. Smith SW, Koppel R. Healthcare information technology's relativity problems: a typology of how patients' physical reality, clinicians' mental models, and healthcare information technology differ. J Am Med Inform Assoc 2014;21(01):117–131 [PubMed: 23800960]
34. Woodfield R, Grant I, Sudlow CLUK Biobank Stroke Outcomes Group UK Biobank Follow-Up and Outcomes Working Group. Accuracy of electronic health record data for identifying stroke cases in large-scale epidemiological studies: a systematic review from the UK Biobank Stroke Outcomes Group. PLoS One 2015;10(10):e0140533 [PubMed: 26496350]
35. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015;350:h1885 [PubMed: 25911572]
36. Schulz S, Jansen L. Formal ontologies in biomedical knowledge representation. Yearb Med Inform 2013;8(9312666):132–146 [PubMed: 23974561]
37. Smith B, Ashburner M, Rosse C, et al. ; OBI Consortium. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25(11):1251–1255 [PubMed: 17989687]
38. Zhang XA, Yates A, Vasilevsky N, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. NPJ Digit Med 2019;2(32):32 [PubMed: 31119199]
39. Bona JP, Prior FW, Zozus MN, Brochhausen M. Enhancing clinical data and clinical research data with biomedical ontologies insights from the knowledge representation perspective. Yearb Med Inform 2019;28(01):140–151 [PubMed: 31419826]
40. Brochhausen M, Bona J, Blobel B. The role of axiomatically-rich ontologies in transforming medical data to knowledge. Stud Health Technol Inform 2018;249:38–49 [PubMed: 29866954]
41. Bowles KH, Potashnik S, Ratcliffe SJ, et al. Conducting research using the electronic health record across multi-hospital systems: semantic harmonization implications for administrators. J Nurs Adm 2013;43(06):355–360 [PubMed: 23708504]
42. Nordo AH, Levaux HP, Becnel LB, et al. Use of EHRs data for clinical research: Historical progress and current applications. Learn Health Syst 2019;3(01):e10076 [PubMed: 31245598]
43. 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;64:333–341 [PubMed: 27989817]
44. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. J Am Med Inform Assoc 2014;21(04):576–577 [PubMed: 24821744]
45. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(02):124–130 [PubMed: 20190053]
46. Murtagh MJ, Blell MT, Butters OW, et al. Better governance, better access: practising responsible data sharing in the META-DAC governance infrastructure. Hum Genomics 2018;12(01):24 [PubMed: 29695297]
47. Meinert E, Alturkistani A, Brindley D, Knight P, Wells G, de Pennington N. Weighing benefits and risks in aspects of security, privacy and adoption of technology in a value-based healthcare system. BMC Med Inform Decis Mak 2018;18(01):100 [PubMed: 30424753]

48. Dankar FK, Ptitsyn A, Dankar SK. The development of large-scale de-identified biomedical databases in the age of genomics-principles and challenges. Hum Genomics 2018;12(01):19 [PubMed: 29636096]
49. Stahl BC, Rainey S, Harris E, Fothergill BT. The role of ethics in data governance of large neuro-ICT projects. J Am Med Inform Assoc 2018;25(08):1099–1107 [PubMed: 29767726]
50. Powles J, Hodson H. Google deepmind and healthcare in an age of algorithms. Health Technol (Berl) 2017;7(04):351–367 [PubMed: 29308344]
51. Roland M Linking physicians' pay to the quality of care–a major experiment in the United Kingdom. N Engl J Med 2004;351(14):1448–1454 [PubMed: 15459308]
52. Roland M, Guthrie B. Quality and outcomes framework: what have we learnt? BMJ 2016;354:i4060 [PubMed: 27492602]
53. United States District Court Northern District of Illinois Eastern Division. MATT DINERSTEIN, individually and on behalf of all others similarly situated, Plaintiff, v. Google, LLC, a Delaware limited liability company, THE UNIVERSITY OF CHICAGO MEDICAL CENTER, an Illinois not-for-profit corporation, and THE UNIVERSITY OF CHICAGO, an Illinois not-for-profit corporation. Accessed June 17, 2021 at: https://www.courtlistener.com/recap/gov.uscourts.ilnd.366172/gov.uscourts.ilnd.366172.85.0.pdf
54. Ohno-Machado L To share or not to share: that is not the question. Sci Transl Med 2012;4(165):165cm15
55. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc 2012;19(05):758–764 [PubMed: 22511014]
56. Ohno-Machado L, Agha Z, Bell DS, et al. ; pSCANNER team. pSCANNER: patient-centered scalable national network for effectiveness research. J Am Med Inform Assoc 2014;21(04):621–626 [PubMed: 24780722]
57. Duan R, Boland MR, Liu Z, 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(03):376–385 [PubMed: 31816040]
58. Duan R, Luo C, Schuemie MJ, et al. Learning from local to global: an efficient distributed algorithm for modeling time-to-event data. J Am Med Inform Assoc 2020;27(07):1028–1036 [PubMed: 32626900]
59. Deng Y, Jiang X, Long Q. Privacy-preserving methods for vertically partitioned incomplete data. Annu Symp Am Med Inform Assoc 2021;2020:348–357
60. Chang C, Deng Y, Jiang X, Long Q. Multiple imputation for analysis of incomplete data in distributed health data networks. Nat Commun 2020;11(01):5467 [PubMed: 33122624]
61. Health Information Privacy. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule. Accessed August 3, 2020. Accessed August 3, 2020 at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de--identification/index.html
62. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care 2012;50:S82–S101 [PubMed: 22692265]
63. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 2019;10(01):3069 [PubMed: 31337762]
64. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010;10(100968545):70 [PubMed: 20678228]
65. Hayden EC. Privacy loophole found in genetic databases. Accessed June 17, 2021 at: https://www.nature.com/articles/nature.2013.12237
66. Pecoraro F, Luzi D, Ricci FL. Designing ETL tools to feed a data warehouse based on electronic healthcare record infrastructure. Stud Health Technol Inform 2015;210:929–933 [PubMed: 25991292]

67. Pecoraro F, Luzi D, Ricci FL. Secondary uses of EHR systems: A feasibility study. Published in: E-Health and Bioengineering Conference (EHB) 21–23 November. 2013; Iasi, Romania; 2013:1–6
68. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit On Translat Bioinforma 2010;2010:1–5
69. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013;46(05):830–836 [PubMed: 23820016]
70. Kroth PJ, Morioka-Douglas N, Veres S, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019;2(08):e199609–e199609 [PubMed: 31418810]
71. National Academy of Medicine, Board on Population Health and Public Health Practice Institute of Medicine. Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Washington, DC: National Academies Press (US); 2015
72. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562(7726):203–209 [PubMed: 30305743]
73. Hamilton CM, Strader LC, Pratt JG, et al. The PhenX Toolkit: get the most from your measures. Am J Epidemiol 2011;174(03):253–260 [PubMed: 21749974]
74. McCarty CA, Berg R, Rottscheit CM, et al. Validation of PhenX measures in the personalized medicine research project for use in gene/environment studies. BMC Med Genomics 2014;7:3 [PubMed: 24423110]
75. Luzi D, Rocco I, Tamburis O, Corso B, Minicuci N, Pecoraro F. Variability in the assessment of children's primary healthcare in 30 European countries. Int J Qual Health Care 2021;33(01):mzab007 [PubMed: 33449077]
76. Deans KJ, Sabihi S, Forrest CB. Learning health systems. Semin Pediatr Surg 2018;27(06):375–378 [PubMed: 30473042]
77. Sarafidis M, Tarousi M, Anastasiou A, et al. Data quality challenges in a learning health system. Stud Health Technol Inform 2020;270:143–147 [PubMed: 32570363]
78. Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. Int J Med Inform 2017;102:71–79 [PubMed: 28495350]
79. Sculley D, Holt G, Golovin D, et al. Hidden technical debt in machine learning systems. Accessed June 17, 2021 at: https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf
80. Mathes T, Buehn S, Prengel P, Pieper D. Registry-based randomized controlled trials merged the strength of randomized controlled trails and observational studies and give rise to more pragmatic trials. J Clin Epidemiol 2018;93:120–127 [PubMed: 28951111]
81. Workman TA. Engaging Patients in Information Sharing and Data Collection: The Role of Patient-Powered Registries and Research Networks. Rockville, MD: Agency for Healthcare Research and Quality; 2013
82. Wozniak L, Soprovich A, Rees S, Johnson ST, Majumdar SR, Johnson JA. Challenges in identifying patients with Type 2 Diabetes for quality-improvement interventions in primary care settings and the importance of valid disease registires. Can J Diabetes 2015;39(Suppl 3):S77–S82 [PubMed: 26145485]
83. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;20(01):117–121 [PubMed: 22955496]
84. Xie S, Himes BE. Approaches to link geospatially varying social, economic, and environmental factors with electronic health record data to better understand asthma exacerbations. AMIA Annu Symp Proc 2018;2018:1561–1570 [PubMed: 30815202]
85. Xie S, Greenblatt R, Levy MZ, Himes BE. Enhancing electronic health record data with geospatial information. AMIA Jt Summits Transl Sci Proc 2017;2017:123–132 [PubMed: 28815121]
86. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the number of health encounters in an electronic health record. Am J Epidemiol 2016;184(11):847–855 [PubMed: 27852603]

87. 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(01):51 [PubMed: 24916006]

88. Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016;23(e1):e20–e27 [PubMed: 26338219]

89. Hubbard RA, Tong J, Duan R, Chen Y. Reducing bias due to outcome misclassification for epidemiologic studies using EHR-derived probabilistic phenotypes. Epidemiology 2020;31(04):542–550 [PubMed: 32282406]

90. Shortreed SM, Cook AJ, Coley RY, Bobb JF, Nelson JC. Challenges and opportunities for using big health care data to advance medical science and public health. Am J Epidemiol 2019;188(05):851–861 [PubMed: 30877288]

91. 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 [PubMed: 29712648]

92. Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: what data are observed and why? EGEMS (Wash DC) 2016;4(01):1203 [PubMed: 27668265]

93. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010;48(06):S106–S113 [PubMed: 20473190]

94. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 2013;20(e2):e206–e211 [PubMed: 24302669]

95. Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn 2003;53:23–69

96. Madhavan S, Bastarache L, Brown JS, et al. Use of electronic health records to support a public health response to the COVID-19 pandemic in the United States: a perspective from 15 academic medical centers. J Am Med Inform Assoc 2021;28(02):393–401 [PubMed: 33260207]

97. Haendel MA, Chute CG, Bennett TD, et al. The national COVID cohort collaborative (N3C): rationale, design, infrastructure, and deployment. J Am Med Inform Assoc 2021;28(03):427–443 [PubMed: 32805036]

98. Brat GA, Weber GM, Gehlenborg N, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med 2020;3:109 [PubMed: 32864472]

99. Burn E, You SC, Sena AG, et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. Nat Commun 2020;11(01):5009 [PubMed: 33024121]

100. Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning learning health system. J Am Med Inform Assoc 2015;22(01):43–50 [PubMed: 25342177]

101. Hutter F, Kotthoff L, Vanschoren J. Automated Machine Learning: Methods, Systems, Challenges. 1st ed. Switzerland: Springer Nature; 2019

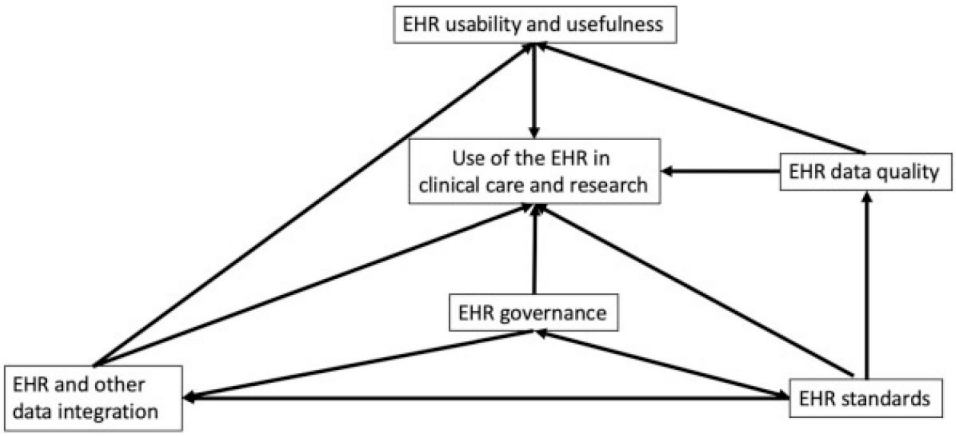

Fig. 1.

Directed acyclic graph illustrating the five entities and activities (facets) that affect the use of the EHR in clinical care and research. The direction of the links indicates the direction of the relationship. For example, standards impact data quality, and standards impact data governance, which in turn impacts data integration. EHR, electronic health record.

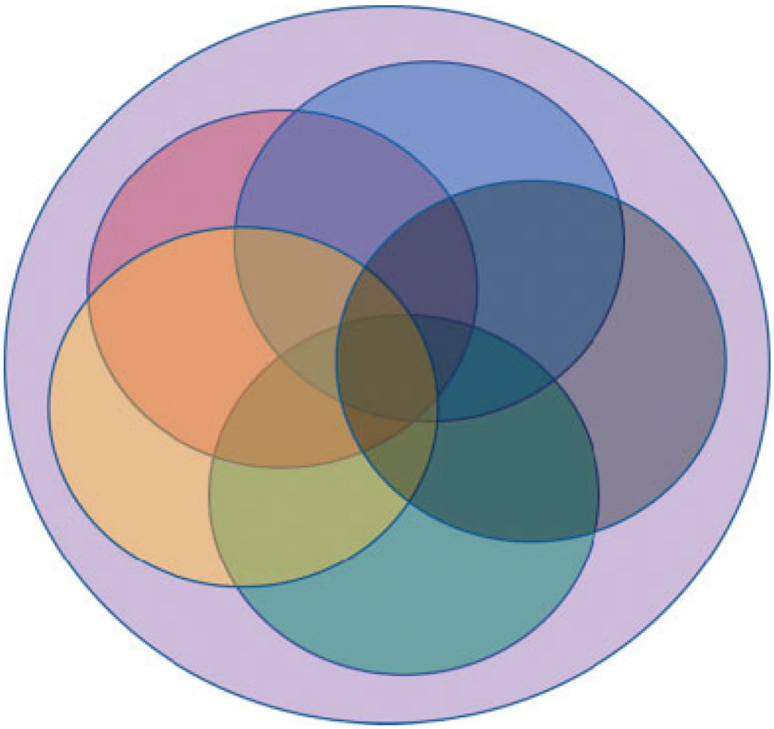

Fig. 2.

Venn diagram illustrating the overlap between the five facets as they relate to each other and the central concept, use of the EHR in clinical care and research. Red: EHR usability and usefulness; green: EHR data quality; blue: EHR standards; yellow: EHR and other data integration; black: EHR governance; violet: use of the EHR in clinical care and research.

Table 1

Matrix illustrating intersections between facets and subfacets

		Usability		Data quality		Data standards				Data governance			Data integration		Clinical care		Clinical research				
		3.1	3.2	4.1	4.2	5.1	5.2	5.3	5.4	6.1	6.2	6.3	7.1	7.2	8.1	8.2	9.1	9.2	9.3	9.4	9.5
Usability	3.1			X	X	X	X		X		X	X	X	X		X	X			X	
	3.2			X	X	X	X	X	X	X	X	X	X	X	X	X	X			X	X
Data quality	4.1	X	X		X	X	X	X	X			X	X		X	X	X	X	X	X	X
	4.2	X	X			X		X	X	X			X	X	X	X	X		X	X	X
Data standards	5.1	X	X	X	X		X	X	X	X	X	X	X	X		X	X				
	5.2	X	X	X		X		X	X		X	X	X	X		X	X	X	X	X	X
	5.3		X	X	X	X	X		X				X	X					X	X	X
	5.4	X	X	X	X	X	X	X		X			X	X		X	X		X	X	XX
Data governance	6.1		X		X	X			X		X	X	X	X		X	X				
	6.2	X	X			X	X			X		X	X	X	X		X	X	X		
	6.3	X	X	X		X	X			X	X		X	X	X		X			X	
Data integration	7.1	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X
	7.2	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X
Clinical care	8.1		X	X	X						X	X	X	X			X	X			X
	8.2	X	X	X	X	X	X		X	X			X	X							
Clinical research	9.1	X	X	X	X	X	X		X	X	X	X	X	X	X			X	X	X	X
	9.2			X			X				X		X	X	X		X				
	9.3			X	X		X	X	X		X		X	X			X	X		X	X
	9.4	X	X	X	X		X	X	X			X	X	X			X	X	X		X
	9.5		X	X	X		X	X	X				X	X	X		X	X	X	X	

Note: “X” indicates specific domains that intersect.