Viewpoint

What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask

Isaac S Kohane[1], MD, PhD; Bruce J Aronow[2], PhD; Paul Avillach[1], MD, PhD; Brett K Beaulieu-Jones[1], PhD; Riccardo Bellazzi[3,4], PhD; Robert L Bradford[5], BSc; Gabriel A Brat[1], MD, MPH; Mario Cannataro[6,7], MS; James J Cimino[8], MD; Noelia García-Barrio[9], MS; Nils Gehlenborg[1], PhD; Marzyeh Ghassemi[10], PhD; Alba Gutiérrez-Sacristán[1], PhD; David A Hanauer[11], MS, MD; John H Holmes[12], PhD; Chuan Hong[1], PhD; Jeffrey G Klann[13,14], PhD; Ne Hooi Will Loh[15], MBBS, FRCA, FFICM, EDIC; Yuan Luo[16], PhD; Kenneth D Mandl[17], MPH, MD; Mohamad Daniar[18], MSIS; Jason H Moore[19], PhD; Shawn N Murphy[1,20], MD, PhD; Antoine Neuraz[21,22], MD; Kee Yuan Ngiam[15], MBBS, MRCS, MMed; Gilbert S Omenn[23], MD, PhD; Nathan Palmer[1], PhD; Lav P Patel[24], MS; Miguel Pedrera-Jiménez[9], MS; Piotr Sliz[17], PhD; Andrew M South[25], MS, MD; Amelia Li Min Tan[1,26], BSc, PhD; Deanne M Taylor[27,28], PhD; Bradley W Taylor[29], MS; Carlo Torti[7], MD; Andrew K Vallejos[29], MS; Kavishwar B Wagholikar[13,14], MBBS, PhD; The Consortium For Clinical Characterization Of COVID-19 By EHR (4CE)[30]; Griffin M Weber[1*], MD, PhD; Tianxi Cai[1*], SCD

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States

[2]Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH, United States

[3]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy

[4]ICS Maugeri, Pavia, Italy

[5]North Carolina Translational and Clinical Sciences Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[6]Data Analytics Research Center, University Magna Graecia of Catanzaro, Catanzaro, Italy

[7]Department of Medical and Surgical Sciences, University Magna Graecia of Catanzaro, Catanzaro, Italy

[8]Informatics Institute, University of Alabama at Birmingham, Birmingham, AL, United States

[9]Department of Informatics, 12 de Octubre University Hospital, Madrid, Spain

[10]Department of Computer Science and Medicine, University of Toronto, Toronto, ON, Canada

[11]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, United States

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

[13]Department of Medicine, Harvard Medical School, Boston, MA, United States

[14]Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA, United States

[15]National University Health Systems, Singapore, Singapore

[16]Department of Preventive Medicine, Northwestern University, Chicago, IL, United States

[17]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States

[18]Clinical Research Informatics, Boston Children's Hospital, Boston, MA, United States

[19]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States

[20]Department of Neurology, Massachusetts General Hospital, Boston, MA, United States

[21]Department of Biomedical Informatics, Necker-Enfant Malades Hospital, Assistance Publique - Hôpitaux de Paris, Paris, France

[22]Centre de Recherche des Cordeliers, INSERM UMRS 1138 Team 22, Université de Paris, Paris, France

[23]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, United States

[24]Department of Internal Medicine, Division of Medical Informatics, University of Kansas Medical Center, Kansas City, KS, United States

[25]Section of Nephrology, Department of Pediatrics, Brenner Children's Hospital, Wake Forest School of Medicine, Winston Salem, NC, United States

[26]Department of Biomedical Informatics, National University of Singapore, Singapore, Singapore

[27]Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA, United States

[28]Department of Pediatrics, Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA, United States

[29]Clinical and Translational Science Institute, Medical College of Wisconsin, Milwaukee, WI, United States

[30]See Acknowledgments

[*]these authors contributed equally

Corresponding Author:
Isaac S Kohane, MD, PhD

Department of Biomedical Informatics
Harvard Medical School
10 Shattuck Street
Boston, MA, 02115
United States
Phone: 1 617 432 3226
Email: isaac_kohane@harvard.edu

Abstract

Coincident with the tsunami of COVID-19–related publications, there has been a surge of studies using real-world data, including those obtained from the electronic health record (EHR). Unfortunately, several of these high-profile publications were retracted because of concerns regarding the soundness and quality of the studies and the EHR data they purported to analyze. These retractions highlight that although a small community of EHR informatics experts can readily identify strengths and flaws in EHR-derived studies, many medical editorial teams and otherwise sophisticated medical readers lack the framework to fully critically appraise these studies. In addition, conventional statistical analyses cannot overcome the need for an understanding of the opportunities and limitations of EHR-derived studies. We distill here from the broader informatics literature six key considerations that are crucial for appraising studies utilizing EHR data: data completeness, data collection and handling (eg, transformation), data type (ie, codified, textual), robustness of methods against EHR variability (within and across institutions, countries, and time), transparency of data and analytic code, and the multidisciplinary approach. These considerations will inform researchers, clinicians, and other stakeholders as to the recommended best practices in reviewing manuscripts, grants, and other outputs from EHR-data derived studies, and thereby promote and foster rigor, quality, and reliability of this rapidly growing field.

(J Med Internet Res 2021;23(3):e22219) doi: 10.2196/22219

KEYWORDS

COVID-19; electronic health records; real-world data; literature; publishing; quality; data quality; reporting standards; reporting checklist; review; statistics

Introduction

What should researchers and clinicians conclude about the recent high-profile retractions of COVID-19 studies based on electronic health record (EHR) data? It is impressive that two publications involving patients with COVID-19, one in The Lancet [1] and the other in the New England Journal of Medicine [2], were determined to be unsound and were retracted in less than 2 months from publication, as these journals' review processes and quality checks are among the most rigorous in the world. Yet, upon closer inspection by those of us familiar with EHR-based research, there were many flaws to these studies involving data quality issues and a lack of transparency that should have been more readily identified during the peer and editorial review process. This is not to say that in-depth statistical analysis might not have eventually uncovered concerns but rather to point out incongruities and anomalies unique to EHR-based studies that should immediately raise concerns to experienced biomedical informaticians, much like an experienced contractor explaining to a homeowner why a competing bid is too good to be true.

In this viewpoint, we present six key questions that are necessary to consider when appraising EHR-based research, especially for research studies investigating the pandemic:

1. How complete are the data?
2. How were the data collected and handled?
3. What were the specific data types?
4. Did the analysis account for EHR variability?
5. Are the data and analytic code transparent?
6. Was the study appropriately multidisciplinary?

In particular, we focus on general aspects of these questions that are crucial to study and data quality and validity of and interpretability of the results and that are broadly applicable to many stakeholders, including researchers and clinicians, in order to optimize the review of submitted manuscripts, published studies, and grant applications containing preliminary data. These desiderata were compiled by the 96 members of the Consortium for Clinical Characterization of COVID-19 by EHR (4CE)—a self-assembled group of collaborating hospitals focused specifically on studying the clinical course of patients with COVID-19 using EHR-based data—most of whom are biomedical informaticians—across 7 countries. 4CE members were invited to contribute their specific key concerns to a shared checklist. This list was then pared down into a less technical list for a more general audience. We excluded those items that are generally considered to be good biostatistical practices (eg, manual review of sample data sets, detecting and understanding outliers [3,4]) to present EHR-specific concerns to a broad biomedical audience. We also excluded recommendations that are contained within the Reporting of Studies Conducted Using Observational Routinely Collected Health Data (RECORD) statement [5,6], which are not specific to EHR-derived data. Finally, we did not focus on the specific limitations of EHR-derived studies, which have been amply documented [7,8], or on the methods to minimize the impact of these limitations,

as this viewpoint is not focused on reviewing specific methodological options for investigators using EHR-derived data, which has been reviewed in detail previously [9-11]. We acknowledge that there are many other criteria that can inform evaluations of EHR-based studies, but we have purposefully limited this discussion to those issues that are most relevant to a general audience, centered on studies investigating the pandemic.

Data Completeness

There are several statistical tests to query data completeness and methods for incorporating missing data [12,13], but here we describe the reasonable expectations for such completeness with knowledge of current, state-of-the-art EHR usage. A publication that is specific about which data were obtained from the EHR (eg, specific laboratory tests or billing codes) is more credible than a study that simply claims it obtained 100% of the EHR data (as did the two recently retracted publications [1,2]). The range of data types from EHRs is extensive and highly varied; each data type requires its own specific quality control and transformations to standard terminologies. For example, laboratory measurements alone can have as many as hundreds of thousands of local codes at a large health care system such as the Veterans Health Administration. In many cases, these data require some level of manual record review to assure data quality and completeness.

Similarly, if a study reports a deidentification procedure, it must describe the details of said procedure. The goals of the deidentification process determine the nature of the deidentification process and the associated regulatory requirements. For example, US hospitals can meet HIPAA (Health Insurance Portability and Accountability Act) standards [14] if they require obfuscation of the counts of patients with rare clinical presentations below a specified prevalence threshold and if they employ date shifting. Knowledge of these methods is essential to analyzing and interpreting the derived data.

Some data types are represented theoretically in the EHR but in practice are only recorded occasionally. For example, standardized codes for smoking history or a family history of specific diseases exist but their underuse is well known. Thus, one cannot assume that the lack of smoking history codes equates to the patient being a nonsmoker. In such scenarios, one must provide an explicit description of the management of missing/null values. Many data elements, such as a complete pulmonary function test, exist in a fragmented form, scattered across different fields in the EHR, and are difficult to extract reliably. In addition, clinical notes allow clinicians greater qualitative expressivity on some of the above values, like smoking history, where they are documented more frequently but not consistently. The quality criteria for reporting narrative content from clinical notes are further addressed below.

Many clinical states are not represented explicitly in the EHR but can be inferred (often referred to as computational phenotypes). When a publication refers to hyperlipidemia, readers should ask themselves whether the hyperlipidemic phenotype is assessed from one or more lipid laboratory tests, billing diagnostic codes, prescription of lipid-lowering medication, or a combination of the above. It is important to document if only structured codes were used or if the phenotype was defined based on information extracted from clinical notes by using natural language processing (NLP) or manual chart review. Either a table describing these phenotypic methods or a reference to a public set of definitions (eg, Phenotype Knowledgebase, PheKB [15]) or a published algorithm with reported accuracy (as seen, for example, in Zhang et al [16] and Ananthakrishnan et al [17]) can provide transparency and precision to these EHR-driven computational phenotypes. The lack of this transparency should be a warning sign. If onset time or temporal trends of clinical events are used as outcomes, it is important to provide sufficient details on how the data were used to derive these outcomes, how granular time was incorporated (eg, by day, 24-hour period, or hour/minute), and to comment on their accuracy, since EHR data are particularly noisy with regards to capturing the timing of events [18,19].

If one uses EHR data to obtain population estimates (eg, prevalence of a complication per 100,000 patients), then additional information should be provided so that readers can determine which subset of patients from that population a given hospital's EHR can capture. For example, if the EHR captures a patient's hospitalization for heart failure, will the EHR also capture the preceding or subsequent outpatient clinic visits related to that hospitalization? With health maintenance organizations, such as Kaiser Permanente, that is much less of a concern, but many hospitals operate in a patchwork system where the patient's data are spread across multiple heterogeneous EHRs that do not necessarily communicate. In our recent COVID-19 study [20], we found many instances in which patients with COVID-19 were transferred from another hospital; unless that other hospital was part of our consortium, it was impossible to have a complete record of their COVID-19 clinical course. It is also important to recognize that a given EHR may not fully capture the clinical course of certain patients, such as those infected with SAR-CoV-2 who have mild symptoms and are discharged home from the emergency room. In these instances, integration of EHR data with data from other sources (eg, primary care providers' offices or nursing homes) may increase the reliability of analysis, although in practice this is rare and such integration methods have to be well documented. EHR systems may also fail to capture acute events that occur outside of the system, especially in the coded data. Leveraging NLP data from the clinical notes can potentially recover partial information if the patient has follow-up visits within that particular system.

Data Collection and Handling

Often the units of measurement and the codes used for data elements like laboratory tests, medications, and diagnoses are not the same across hospitals and may even differ within the same health care system or change over time. Single analytic concepts (eg, the troponin T test) can balloon into dozens of local codes at each hospital, since these tests may be performed at different diagnostic laboratories, each with its own distinct codes or with different technologies over time. Therefore, they have to be "harmonized," or mapped, to agreed-upon standard terminologies and scales [21]. Even when they are the same,

their meaning can differ based on population or practice differences (eg, which sensitive troponin test is used or which reference range defines a test result being normal, or in children rather than in adults, whose normative values often change across the age range) [7]. In both instances, readers should expect that the specific procedures for harmonization or site-specific semantic alignment are described adequately in the Methods section (or via supplementary materials). A summary of this process can become increasingly complex within the usual confines of a Methods section for multisite and international studies where, by necessity, the site-by-site variability is high.

Data Type

There are large methodological divides and divergent ethical challenges between codified data (eg, discrete laboratory values such as serum glucose) and narrative text (eg, discharge summary) from which characterizations are obtained using NLP. While both data types have their own limitations, methods that incorporate both can greatly improve the sensitivity and/or specificity of the clinical characterizations and phenotyping of a group of patients. For example, signs and symptoms are often not codified discreetly or consistently (eg, not entered into the EHR's Problem List) but are written in the clinical notes. Similarly, outpatient medication documentation in clinical notes does not necessarily represent accurately the medications that the patient is actually taking, but prescriptions entered into the EHR may. Combining both codified and NLP data can substantially improve sensitivity and/or specificity and ideally one should always use this complementarity [22-24]. For example, only about 10% of pregnant women with suicide ideation have related codes and vast majority of the cases are only documented in the notes [25]. However, the ability to extract NLP data and the accuracy of those data may be limited by each institution's informatics infrastructure and expertise as well as local institutional review board (IRB) constraints. Furthermore, NLP application to clinical narrative text is relatively new and more prone to large variability in the quality of the obtained characterizations. Particularly in countries with different languages, the NLP techniques and their performance may vary widely. For this reason, readers should expect a reference to the specific NLP methods used and their performance characteristics on data of the sort that the study collected and analyzed. For example, if someone describes the use of an NLP approach on discharge summaries in intensive care units in Italy, but the provided citation was validated only for use in outpatient notes written in English, readers can be legitimately concerned about the accuracy and validity of the patient characterizations in that study. Furthermore, if a study claims very high accuracy, readers should expect a report (or citation of a report) that shows an expert review of the NLP method validated against a representative sample confirming the claimed performance.

Robustness Against EHR Variability

Beyond any variation in human biology across countries and continents, different styles of practice, and how different reimbursement schemes influence styles of practice and use of EHRs, have a very large impact on the nature of EHR data. Therefore, a multinational study should at least acknowledge these differences as a limitation or explicitly attempt to account for them in the analyses. For example, in COVID-19–related research, it has become increasingly apparent that there is an association between patient race/ethnicity and their risk for acquisition of and complications from COVID-19. However, this association is much less detectable in EHR data, as, for example, it is mostly invisible in data from Europe because several countries forbid collecting self-reported race in the EHR. Even in the United States, the coding of different ethnicities or multiracial identification is not standardized. In addition, some countries have far more comprehensive primary care EHR data sharing, whereas others (like the United States) cannot aggregate data systematically and consistently across major health care centers.

Transparency

In order to ensure patients' rights to privacy, patient-level data can rarely be shared outside an institution. In many EHR-driven studies, the code to extract data from a source EHR can be protected by confidentiality agreements with the EHR vendor and is thus difficult to share. Nonetheless, the code or algorithm for creating the variables used for analyses should be provided even if the detailed data extraction procedures are not shared because of commercial restrictions. Running the code on synthetic data sets that follow a standard data model can demonstrate code functionality and facilitate code reuse [26]. The code used to conduct statistical analyses and create visualizations—after data extraction—should also be shared in public repositories to enable other researchers to follow each step of the analysis and provide further transparency. While there are significant challenges to sharing patient-level data, one can share intermediate results and aggregate distributions to increase transparency and understand between-institution differences [27]. One should archive the data used for analyses, along with the associated data extraction codes, at the local institution to ensure reproducibility. Authors should also make the deidentified data available—either publicly in a repository or by request. While only a small fraction of readers typically look at the code, whether referenced on a file server or shared as supplementary methods, the availability of the code provides reassurance and validation that the study utilized proper methodologies.

Multidisciplinary Approach

There may come a time when data can be aggregated automatically from multiple EHR environments to answer a particular question without relying on a human to understand the particular idiosyncrasies of each institution's data and EHR system. Until that day, effective EHR data set analysis requires collaboration with clinicians and scientists who have knowledge of the diseases being studied and the practices of their particular health care systems; informaticians with experience in the underlying structures of biomedical record repositories at their own institutions and the characteristics of their data; data

harmonization experts to help with data transformation, standardization, integration, and computability; statisticians and epidemiologists well versed in the limitations and opportunities of EHR data sets and related sources of potential bias; machine learning experts; and at least one expert in regulatory and ethical standards. Data provenance records should already exist to ensure compliance with privacy standards, so that authors can readily point to these processes and reference institutional officials who grant data access similarly to IRBs. In our experience, we often have an interdisciplinary team participate in the process of establishing the research question and study design, defining the data elements, and determining what analyses can be performed given the available data. It is also important that people with complementary skills work together to review and interpret the results [28]. Each of these steps is a major contribution deserving of authorship. Just as a population genetics study reporting across countries often has dozens of authors, so do we expect multihospital EHR-driven studies to acknowledge and name the individuals as authors and in doing so provide accountability for the dozens of procedures, checks, and balances necessary for the reliable extraction of EHR patient data. Consequently, contribution statements should list explicitly the responsibilities of each author with regard to study conceptualization and design, data extraction, data harmonization, data integration, data analysis, results interpretation, and regulatory and ethical oversight. Additionally, although reputation is sometimes overvalued, having *no* reputation or at least a track record of appropriate success should trigger greater attention to documenting the process to reach the same level of trust. Unlike a mathematical proof, simple inspection of the data may be insufficient and will become increasingly so in the era of data generated by machine learning algorithms purposefully built for the task of conditioning data to appear real. Trust and accountability become essential companions to transparency and clarity during the EHR analytic process.

Conclusion

Similar to publications from the early days of the genomic revolution, which initially included extensive sections on DNA sequencing validation, methods, reagents, and conditions that became progressively briefer as trust was built and the methods commoditized, comprehensively and transparently reported methods of EHR data extraction and transformation are at least as important as subsequent statistical analysis and interpretation. We need to be open and transparent about the inherent limitations of the data and the analyses. We should also acknowledge alternative interpretations of the results (eg, outlier prescribing practices in one country that confound the apparent effects of that drug in that country). Extra caution is also needed in how we draw causal inferences from EHR data, especially given the noisiness and incompleteness of the data in addition to several sources of bias, though application of a causal model framework and specific causal inference methods may help mitigate some of these concerns. The recommendations we have outlined here (see Table 1 for our 12-item checklist) do not substitute for a durable research infrastructure that would enable tracking EHR data provenance along explicit source, ownership, and data protocols, which would allow for rigorous and routine quality assurance in the use of EHR data [29].

Table 1. 12-item checklist to assess electronic health record (EHR) data–driven studies.

Item	Reassuring	Concerning
Defining study cohort/data extraction	Reporting the precise definition of the domains and/or subsets of EHR data extracted for the study cohort and the information system sources	100% of the EHR said to be extracted or no specification of which subsets of the EHR data were obtained
Deidentification	Specific deidentification algorithm documented with acknowledgment of analytic consequences/limitations	Only a statement that deidentification was performed
Defining clinical variables/data type–specific omissions/limitations	For data types represented poorly in EHR codified data, either NLP[a] is deployed on the EHR clinical notes or additional data sources (eg, self-reported questionnaires) are used. Procedures to deal with missing values should also be made explicit	Referencing data types like family/social history without explaining how they are obtained through NLP or exceptional codified data practice
Phenotypic transparency	Computational phenotypes that are more than just a specific native EHR variable (eg, hyperlipidemia vs a specific LDL[b] measurement) are either defined in the study or a citation is given to algorithmic phenotype definitions	Clinical phenotypes are used in the study without specifying how they were derived from the EHR data
Generalizing EHR findings to the population/population denominator	Study heavily cautions on using prevalence/incidence estimates from the EHR data or refers to empirical estimates on how much of a patient's entire health care is captured in that particular EHR	Direct estimates of prevalence or incidence from EHR frequencies without justifying that generalization
Data collection	Clinical forms or data models implemented in health care information systems are shared or clearly described. This includes the coding systems used	Mention structured data without specifying the clinical forms or data models. Mention coded data without mentioning coding systems
Data transformation/harmonization	Data transformation process shared or clear description of which methods were used to harmonize data to a standardized terminology, scale units, and account for different local usage	Mention of harmonization methods without specifying which ones and what problems were identified and addressed/overcome
Textual vs codified data	If textual data are used in the study, then specification of which clinical notes, in what language, with which NLP algorithm with either an explanation of or a citation to that algorithm's validation, sensitivity, and specificity for comparable data	Harmonization efforts for codified and textual data treated as if they are the same process. Lack of specificity in describing the NLP algorithm and performance
Manual coding of data	Qualifications of coders described, formal coding criteria described or at least mentioned, intercoder reliability measured and reported	No description of process for turning text or nonstandard coded data into standard coded data; use of crowd-sourced coders (eg, graduate students or Mechanical Turk) without mention of quality assurance processes
Regional and global variation	A study describes how they adjust for (or exclude) differences that are due to variation in practice, regulation, and clinical documentation through the EHR from site to site	A study says they adjusted for regional or country differences in practice or EHR documentation but do not describe how they do it
Sharing analytic code	Analytic code is deposited in a public repository or study-specific public website	Code is not shared or only "shared on demand"
Acknowledge a multidisciplinary team	Authorships for all parts of the extraction-through-analysis pipeline with precision as to each contribution	Health care system sources not named or local health care system site collaborators not named

[a]NLP: natural language processing.

[b]LDL: low-density lipoprotein.

Finally, in crises such as the COVID-19 pandemic, we need to recognize that many studies can contribute to our understanding of what is happening to our patients and how our practices might affect patient outcomes. Overly generalized conclusions will likely strain the boundaries of what can be reasonably inferred from the kinds of data currently obtained through EHRs. Recommendations that flow from overly broad claims may irreversibly harm stakeholders, including patients and clinicians. Increased reader awareness of EHR-derived data quality indicators is crucial in critically appraising EHR-driven studies and to prevent harm from misleading studies, which will ensure sustainable quality in this rapidly growing field.

Acknowledgments

The members of the Consortium for Clinical Characterization of COVID-19 By EHR (4CE) are as follows: Adem Albayrak, Danilo F Amendola, Li LLJ Anthony, Bruce J Aronow, Andrew Atz, Paul Avillach, Brett K Beaulieu-Jones, Douglas S Bell, Antonio Bellasi, Riccardo Bellazzi, Vincent Benoit, Michele Beraghi, José Luis Bernal Sobrino, Mélodie Bernaux, Romain Bey,

Alvar Blanco Martínez, Martin Boeker, Clara-Lea Bonzel, John Booth, Silvano Bosari, Florence T Bourgeois, Robert L Bradford, Gabriel A Brat, Stéphane Bréant, Mauro Bucalo, Anita Burgun, Tianxi Cai, Mario Cannataro, Aize Cao, Charlotte Caucheteux, Julien Champ, Luca Chiovato, James J Cimino, Tiago K Colicchio, Sylvie Cormont, Sébastien Cossin, Jean Craig, Juan Luis Cruz Bermúdez, Arianna Dagliati, Mohamad Daniar, Christel Daniel, Anahita Davoudi, Batsal Devkota, Julien Dubiel, Scott L DuVall, Loic Esteve, Shirley Fan, Robert W Follett, Paula SA Gaiolla, Thomas Ganslandt, Noelia García Barrio, Nils Gehlenborg, Alon Geva, Tobias Gradinger, Alexandre Gramfort, Romain Griffier, Nicolas Griffon, Olivier Grisel, Alba Gutiérrez-Sacristán, David A Hanauer, Christian Haverkamp, Martin Hilka, John H Holmes, Chuan Hong, Petar Horki, Meghan R Hutch, Richard Issitt, Anne Sophie Jannot, Vianney Jouhet, Mark S Keller, Katie Kirchoff, Jeffrey G Klann, Isaac S Kohane, Ian D Krantz, Detlef Kraska, Ashok K Krishnamurthy, Sehi L'Yi, Trang T Le, Judith Leblanc, Guillaume Lemaitre, Leslie Lenert, Damien Leprovost, Molei Liu, Ne Hooi Will Loh, Yuan Luo, Kristine E Lynch, Sadiqa Mahmood, Sarah Maidlow, Alberto Malovini, Kenneth D Mandl, Chengsheng Mao, Patricia Martel, Aaron J Masino, Michael E Matheny, Thomas Maulhardt, Michael T McDuffie, Arthur Mensch, Marcos F Minicucci, Bertrand Moal, Jason H Moore, Jeffrey S Morris, Michele Morris, Karyn L Moshal, Sajad Mousavi, Danielle L Mowery, Douglas A Murad, Shawn N Murphy, Kee Yuan Ngiam, Jihad Obeid, Marina P Okoshi, Karen L Olson, Gilbert S Omenn, Nina Orlova, Brian D Ostasiewski, Nathan P Palmer, Nicolas Paris, Lav P Patel, Miguel Pedrera Jimenez, Hans U Prokosch, Robson A Prudente, Rachel B Ramoni, Maryna Raskin, Siegbert Rieg, Gustavo Roig Domínguez, Elisa Salamanca, Malarkodi J Samayamuthu, Arnaud Sandrin, Emily Schiver, Juergen Schuettler, Luigia Scudeller, Neil Sebire, Pablo Serrano Balazote, Patricia Serre, Arnaud Serret-Larmande, Domenick Silvio, Piotr Sliz, Jiyeon Son, Andrew M South, Anastasia Spiridou, Amelia LM Tan, Bryce WQ Tan, Byorn WL Tan, Suzana E Tanni, Deanne M Taylor, Valentina Tibollo, Patric Tippmann, Andrew K Vallejos, Gael Varoquaux, Jill-Jênn Vie, Shyam Visweswaran, Kavishwar B Wagholikar, Lemuel R Waitman, Demian Wassermann, Griffin M Weber, Yuan William, Zongqi Xia, Alberto Zambelli, Aldo Carmona, Charles Sonday, and James Balshi.

Authors' Contributions

ISK led the 4CE international consortium, conceived and designed the study, and drafted the manuscript. TC led 4CE analytics strategies and made contributions to the study design and drafting of the manuscript. JJC contributed a validation strategy and made edits to the manuscript. NG-B was responsible for data extraction and transformation to 4CE format and quality control of the results and made internal contributions. NG led 4CE visualization strategies and made contributions/edits to the manuscript. JGK contributed to the 4CE validation strategy and data submission strategies and made edits to the manuscript. KDM made contributions to the text and framework and made edits to the manuscript. DM was involved in data extraction and transformation to 4CE format. SNM led 4CE data validation strategies and made contributions/edits to the manuscript. GSO made contributions to strategy and edits to the manuscript. NP contributed to 4CE data analysis, aggregation, and quality control. KBW contributed to validation strategies and made edits to the manuscript. BJA, PA, BKB-J, RB, RLB, GAB, MC, MG, AG-S, DAH, JHH, CH, NHW, YL, JHM, AN, KYN, LPP, MP-J, PS, AMS, ALMT, DMT, BMT, CT, AKV, and GMW made contributions/edits to the manuscript.

Conflicts of Interest

RB and AM are shareholders of Biomeris srl. GSO is affiliated with BoD, Galectin Therapeutics, Angion Biomedica, and Amesite, Inc. DMT consulted on a legal matter for AstraZeneca last year.

References

1. Mehra MR, Desai SS, Ruschitzka F, Patel AN. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. The Lancet. May 2020. [FREE Full text] [doi: 10.1016/S0140-6736(20)31180-6]
2. Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med. Jun 18, 2020;382(25):e102. [doi: 10.1056/nejmoa2007621]
3. Cox D, Donnelly C. Principles of Applied Statistics. Cambridge, UK. Cambridge University Press; 2011.
4. Eriksson L, Byrne T, Johansson E, Trygg J, Vikström C. Multi- and Megavariate Data Analysis Basic Principles and Applications. Malmo, Sweden. Umetrics Academy; 2013.
5. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. Oct 6, 2015;12(10):e1001885. [FREE Full text] [doi: 10.1371/journal.pmed.1001885] [Medline: 26440803]
6. Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ. Nov 14, 2018;363:k3532. [FREE Full text] [doi: 10.1136/bmj.k3532] [Medline: 30429167]
7. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. Medical Care. 2013;51:S30-S37. [doi: 10.1097/mlr.0b013e31829b1dbd]

8. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res. May 29, 2018;20(5):e185. [FREE Full text] [doi: 10.2196/jmir.9134] [Medline: 29844010]
9. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). Sep 11, 2016;4(1):1244. [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
10. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. Jan 01, 2013;20(1):144-151. [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
11. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. Annu Rev Public Health. Mar 18, 2016;37(1):61-81. [FREE Full text] [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
12. Capocaccia R, De Angelis R. Estimating the completeness of prevalence based on cancer registry data. Statist Med. Feb 28, 1997;16(4):425-440. [doi: 10.1002/(sici)1097-0258(19970228)16:4<425::aid-sim414>3.0.co;2-z]
13. Smirnov VB. Earthquake catalogs: Evaluation of data completeness. Volc Seis. 1998;19:497-510. [FREE Full text]
14. Methods for De-identification of PHI. Office for Civil Rights. Nov 6, 2015. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [accessed 2020-06-16]
15. Kirby J, Speltz P, Rasmussen L, Basford M, Gottesman O, Peissig P, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. Nov 2016;23(6):1046-1052. [FREE Full text] [doi: 10.1093/jamia/ocv202] [Medline: 27026615]
16. Zhang J, Can A, Lai PMR, Mukundan S, Castro VM, Dligach D, et al. Age and morphology of posterior communicating artery aneurysms. Sci Rep. Jul 14, 2020;10(1):11545. [FREE Full text] [doi: 10.1038/s41598-020-68276-9] [Medline: 32665589]
17. Ananthakrishnan AN, Cagan A, Cai T, Gainer VS, Shaw SY, Churchill S, et al. Statin Use Is Associated With Reduced Risk of Colorectal Cancer in Patients With Inflammatory Bowel Diseases. Clin Gastroenterol Hepatol. Jul 2016;14(7):973-979. [FREE Full text] [doi: 10.1016/j.cgh.2016.02.017] [Medline: 26905907]
18. Uno H, Ritzwoller DP, Cronin AM, Carroll NM, Hornbrook MC, Hassett MJ. Determining the Time of Cancer Recurrence Using Claims or Electronic Medical Record Data. JCO Clinical Cancer Informatics. Dec 2018;(2):1-10. [doi: 10.1200/cci.17.00163]
19. Liu C, Wang F, Hu J, Xiong H. Temporal Phenotyping from Longitudinal Electronic Health Records: A Graph Based Framework. New York, NY. Association for Computing Machinery; 2015. Presented at: KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data; August 2015:705-714; Sydney, NSW, Australia. [doi: 10.1145/2783258.2783352]
20. Brat G, Weber G, Gehlenborg N, Avillach P, Palmer N, Chiovato L, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med. 2020;3:109. [FREE Full text] [doi: 10.1038/s41746-020-00308-0] [Medline: 32864472]
21. Klann J, Abend A, Raghavan V, Mandl K, Murphy S. Data interchange using i2b2. J Am Med Inform Assoc. Sep 2016;23(5):909-915. [FREE Full text] [doi: 10.1093/jamia/ocv188] [Medline: 26911824]
22. Ananthakrishnan AN, Cai T, Savova G, Cheng S, Chen P, Perez RG, et al. Improving Case Definition of Crohn s Disease and Ulcerative Colitis in Electronic Medical Records Using Natural Language Processing. Inflammatory Bowel Diseases. 2013;19(7):1411-1420. [doi: 10.1097/mib.0b013e31828133fd]
23. Ning W, Chan S, Beam A, Yu M, Geva A, Liao K, et al. Feature extraction for phenotyping from semantic and knowledge resources. J Biomed Inform. Mar 2019;91:103122. [FREE Full text] [doi: 10.1016/j.jbi.2019.103122] [Medline: 30738949]
24. Zhang Y, Cai T, Yu S, Cho K, Hong C, Sun J, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc. Dec 20, 2019;14(12):3426-3444. [FREE Full text] [doi: 10.1038/s41596-019-0227-6] [Medline: 31748751]
25. Zhong Q, Karlson EW, Gelaye B, Finan S, Avillach P, Smoller JW, et al. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. BMC Med Inform Decis Mak. May 29, 2018;18(1):30. [FREE Full text] [doi: 10.1186/s12911-018-0617-7] [Medline: 29843698]
26. Morin A, Urban J, Adams PD, Foster I, Sali A, Baker D, et al. Research priorities. Shining light into black boxes. Science. Apr 13, 2012;336(6078):159-160. [FREE Full text] [doi: 10.1126/science.1218263] [Medline: 22499926]
27. Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. Nat Biotechnol. Apr 13, 2017;35(4):342-346. [FREE Full text] [doi: 10.1038/nbt.3780] [Medline: 28288103]
28. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. Apr 24, 2015;350(apr24 11):h1885-h1885. [FREE Full text] [doi: 10.1136/bmj.h1885] [Medline: 25911572]
29. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data: a transnational perspective. Int J Med Inform. Jan 2013;82(1):1-9. [doi: 10.1016/j.ijmedinf.2012.11.003] [Medline: 23182430]

Abbreviations

4CE: Consortium for Clinical Characterization of COVID-19 by EHR
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
RECORD: Reporting of Studies Conducted Using Observational Routinely Collected Health Data
NLP: natural language processing
IRB: institutional review board
PheKB: Phenotype Knowledgebase

Edited by R Kukafka; submitted 13.07.20; peer-reviewed by N Delvaux, M Adly, P Harris, A Adly, A Adly, J Li, L Genaro; comments to author 04.08.20; revised version received 14.09.20; accepted 10.01.21; published 02.03.21

Please cite as:
Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, Brat GA, Cannataro M, Cimino JJ, García-Barrio N, Gehlenborg N, Ghassemi M, Gutiérrez-Sacristán A, Hanauer DA, Holmes JH, Hong C, Klann JG, Loh NHW, Luo Y, Mandl KD, Daniar M, Moore JH, Murphy SN, Neuraz A, Ngiam KY, Omenn GS, Palmer N, Patel LP, Pedrera-Jiménez M, Sliz P, South AM, Tan ALM, Taylor DM, Taylor BW, Torti C, Vallejos AK, Wagholikar KB, The Consortium For Clinical Characterization Of COVID-19 By EHR (4CE), Weber GM, Cai T
What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask
J Med Internet Res 2021;23(3):e22219
URL: https://www.jmir.org/2021/3/e22219
doi: 10.2196/22219
PMID: 33600347