



Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview

Ban Al-Sahab¹ · Alan Leviton^{2,3} · Tobias Loddenkemper^{2,3} · Nigel Paneth^{4,5} · Bo Zhang^{3,6,7}

Received: 30 June 2023 / Revised: 5 September 2023 / Accepted: 7 November 2023 /

Published online: 14 November 2023

© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

Abstract

Electronic Health Records (EHR) are increasingly being perceived as a unique source of data for clinical research as they provide unprecedentedly large volumes of real-time data from real-world settings. In this review of the secondary uses of EHR, we identify the anticipated breadth of opportunities, pointing out the data deficiencies and potential biases that are likely to limit the search for true causal relationships. This paper provides a comprehensive overview of the types of biases that arise along the pathways that generate real-world evidence and the sources of these biases. We distinguish between two levels in the production of EHR data where biases are likely to arise: (i) at the healthcare system level, where the principal source of bias resides in access to, and provision of, medical care, and in the acquisition and documentation of medical and administrative data; and (ii) at the research level, where biases arise from the processes of extracting, analyzing, and interpreting these data. Due to the plethora of biases, mainly in the form of selection and information bias, we conclude with advising extreme caution about making causal inferences based on secondary uses of EHRs.

Keywords Electronic Health Records · Real World Data · Real World Evidence · Bias · Study Validity

1 Introduction

Fueled by evolving federal regulation [1–3] and the Covid-19 pandemic [3–5], real-world data (RWD), largely in the form of Electronic Health Records (EHRs) [6], are increasingly being used to generate real-world evidence (RWE). While primarily intended for clinical and administrative use, EHRs contain massive amounts of information about patients’ “*medical history including, diagnoses, treatment plans,*

immunization dates, allergies, radiology images, pharmacy records, and laboratory and test results” [7] that are available for secondary use in research [1]. Unlike randomized clinical trials (RCT) which are largely conducted in specialized research environments with highly selected participants, EHRs reflect health and disease in the general population and the realities of clinical practice [8–10]. EHRs can also provide RWE more quickly and at a lower cost than RCTs [8].

We acknowledge that collections of EHRs can identify trends, [11–15] assist in creating clinical decision supports, [14, 16] and provide valuable drug-post-marketing information, [10, 17–21] associations, [11–13] and opportunities for improvement [14–16]. On the other hand, we advise against their use for making decisions about causation, unless unusual circumstances apply, such as having uncovered very large relative risks in settings where bias and confounding are minimal. We are most concerned with how faulty or incomplete data in the EHR might result in errors made in drawing inferences about the relationship between characteristics of the patient or treatment and the occurrence, severity, and course of the patient’s disease or response to a treatment. We specifically focus in this paper on the challenges that limit the usefulness of RWE generated from observational studies that are based on EHR data. We provide a comprehensive overview of the sources and types of biases in EHRs research, offer words of caution to readers, and outline potential remedies.

In this paper we identify two sets of processes that are likely to be sources of biases that arise along the pathway of generating RWE from EHRs. The first process is at the healthcare system level, where the selection of patients and healthcare interventions, and imperfect data collection create biases. Specifically, in this setting, EHR data might poorly reflect the actual experiences of patients with the conditions studied. The second process is at the research level including the design of the study, the extraction, analysis, and interpretation of the data, which can distort inferences derived from RWD. Figure 1 illustrates these two sets of processes with a breakdown of potential sources of bias associated with each one. These sources of bias also represent the flow of RWD from its initial creation to its transformation into RWE, a concept first identified by Verheij et al. 2018 [17].

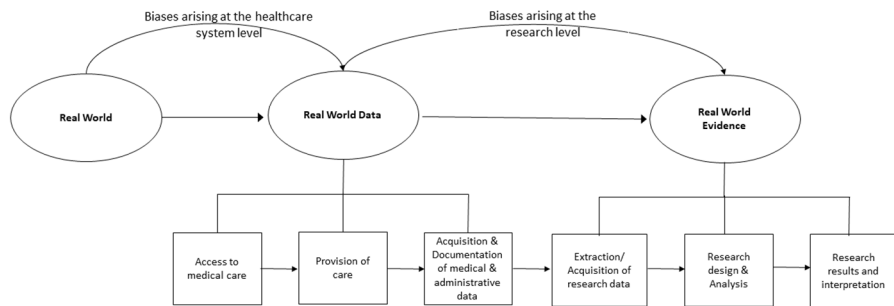


Fig. 1 Pathway of generating RWE from EHR data

2 Biases Arising at the Healthcare System Level

Research based on EHR data is subject to selection and information biases, that are inherent in the healthcare system's patient population and its EHR format. Selection bias, defined as “*systematic differences in characteristics between those who are selected and those who are not*” [18], mainly arises from a lack of representativeness because the population captured in the EHR rarely fully represents the population that is the source of the EHR population. Further, selection bias arises from data missingness, either from missing visits or from missing information within visits. Information bias, on the other hand, which is “*a flaw in measuring exposure or outcome that results in differential quality of information between compared groups*” [18], mainly arises due to misclassification of the clinical picture (EHR discontinuity, missing/incomplete data, and inability of the EHR system to capture the patient's true health status), measurement errors, and variability in data collection methods.

3 Access to Medical Care

3.1 EHR Representativeness

Unlike the general population, all individuals represented in the EHR system have sought and obtained medical care from a specific set of providers [19]. Healthy individuals and patients with milder diseases, without medical coverage, or who use other practitioners are unlikely to be represented in a specified EHR system [20, 21]. The geographic “*catchment*” area of the EHR is the first restriction of the population, but social, demographic, and economic factors further determine the patient's enrollment in any given EHR [5]. Additional constraints include distance to the healthcare provider, number of available healthcare systems, type of healthcare system [22, 23], and health insurance plan [22, 24]. EHR systems will preferentially include women, elderly, whites, the more educated, and others more likely to seek medical care, especially primary care [21].

Lack of representativeness is one of the major challenges of EHR data [5, 19], introducing selection bias when study inferences are generalized to broader target populations [19, 24–26]. Unlike population-based studies, which also include non-recipients of medical care, and where participant selection is determined by the study sampling plan, enrollment in EHR is largely driven by the individual patient [19]. Because the mechanisms that drive patients to interact with the healthcare system are incompletely known, analytic mechanisms cannot fully address and control for them [19, 27].

3.2 EHR Discontinuity

Patients whose medical information is captured in EHRs are considered members of “*open cohorts*,” where they enter and leave the record system during the period of observation [28, 29]. Patients can drop in or out of an EHR system to seek specialty care elsewhere, to return to a primary healthcare provider [19], or drop out of

the system entirely because of changes in their disease characteristics or insurance coverage [29].

Patients are referred to as “*censored*” for the time when they are not being followed. The absence of patient information *prior* to their entry into the EHR is referred to as *left* censoring. Absence *after* their departure from the EHR is referred to as *right* censoring [5, 29]. Absence of information for a period of time with information available both before and after that period is referred to as *interval* censoring. The movement of patients across provider systems can result in all three forms of censoring in EHR research.

Censoring can cause structural missingness in EHRs [30], leading to an incomplete picture of the history of disease in a given patient. Components of the diagnosis, progression, management, and treatment of disease, as well as the time of their occurrence, are often missing. These absences can lead to information bias and/or selection bias depending on how the researcher deals with the missingness. Removing patients with censoring is one solution but may increase the problem of selection bias [31], even with use of advanced statistical techniques to address biases [19, 22, 27, 32]. Information bias, on the other hand, arises when these patients are included but outcomes are misclassified due to the EHR’s inability to capture the true history of the disease [19].

4 Provision of Care

Whether a clinical event is recorded in the EHR is dependent on policies, practices, referral and reimbursement systems, and professional guidelines that differ across healthcare systems, creating “*healthcare system bias*” [17]. These factors affect the initiation, frequency, content, and documentation of clinical encounters recorded in the EHR [17]. For example, professional guidelines and reimbursement systems promote blood pressure readings at most in-person visits in some countries, but only if required by clinical conditions in others, creating *selection bias* for this measure [17].

Patient encounters can take place in inpatient or outpatient settings and can represent primary, specialty, or emergent care, depending on the services and settings provided by the healthcare system. These settings, as well as the practice workload, shape the nature and intensity of data recorded in the EHR [22]. Provision of medical services might also be influenced by the patient’s type of insurance [4]. Understanding the coding processes in the EHR system is essential to minimize *information bias* [17], since insurance type often dictates billing codes used to establish diagnoses and treatments in EHRs [4, 5, 23].

Moreover, the 21st Century Cures Act [33] requires all medical-care providers to offer patients access to their EHR information, which may influence the recording of data thought to carry a stigma or might stimulate medical litigation [17, 34].

5 Acquisition and Documentation of Medical and Administrative Data

5.1 Data Collection and Measurement

EHR data are composed of structured fields (smart forms and clinical templates), unstructured data (clinical notes, free text), and peripheral documents (imaging data, pathology reports) [10]. Clinical information is often obtained by examination all of these EHR sections [35]. Due to the variability in data collection methods, EHR data are subject to misclassification bias and/or measurement errors. Recent papers have recommended statistical techniques for minimizing such biases in EHR research [19, 27, 36–38].

5.1.1 Structured Data

The International Classification of Diseases (ICD) is the main diagnostic coding scheme used across the US healthcare system [4, 39], and researchers need to consider any changes to the coding system or provider's terminology over the years, as occurred in 2015 when the US healthcare system transitioned from ICD-9 to ICD-10 [4, 5]. The presence of an ICD code in the EHR indicates the existence of a disease, but its absence does not assure the absence of the disease. ICD codes in the EHR thus have high specificity but low sensitivity [28]. When the diagnosis is not obvious, a patient can also have a series of “*rule-out*” ICD codes for testing and follow-up until the final correct diagnosis is made, and these rule out codes can be mistaken for actual diagnoses [39].

Disease status may be incompletely estimated from diagnostic and billing codes, while information on other structured data may provide a fuller picture of disease status [35]. Although a structured item, recording medication, is also not straightforward. The same chemical formulation of a medication can be described under the rubric of several brands or generic names [39], and in extended or immediate release variations. While information on ordered drugs is usually recorded in the EHR, information on whether the patient actually obtained and took the prescribed medication at the required dosage and timing is rarely available in the EHR. Many EHRs are not linked to pharmacy dispensing information systems [39].

Laboratory measures and vital signs are also some of the readily available structured data in EHR but its use in research has its own challenges [39]. The main standard coding systems for laboratory tests and results are the Logical Observation Identifiers Names and Codes (LOINC), the Systematized Nomenclature of Medicine (SNOMED), and the Current Procedural Terminology (CPT) [40]. Unfortunately, implementation is still not universal [1] and many clinical laboratories use local coding systems [40]. Similarly, the coding of vital signs is generally made without standardized approaches, resulting in inconsistency. According to the working group of the National Heart, Lung, and Blood Institute, blood pressure measurement is still sub-optimal in clinical practice and deviates from the recommended guidelines [41]. Measures that are self-reported can also differ from those obtained by medical professionals resulting in additional sources of variance [40, 42].

Substantial variability in the quality of data recorded in EHR has also been attributed to the software packages [17, 43]. The fields are at times flawed either because they are too broad or require a level of unnecessary details or demand a response/value for an inapplicable field [4]. EHR software can also change over time to reflect new emerging technologies or new changes in coding or billing practices [4, 44].

5.1.2 Unstructured Data

Unstructured or semi-structured data are any form of data available at EHR that does not conform to a pre-designed organized structure [39]. Clinician notes, for example, can provide important information not present in structured fields, such as changes in symptom severity and side effects of medication [5, 7, 9]. Artificial intelligence approaches, including natural language processing (NLP) and machine learning (ML) may be able to transform unstructured notes into useful data [7].

In addition, patients now have electronic opportunities to record self-reported events (e.g. asthma attacks, seizures), episodes of pain, and periodic assessments of “*quality of life*” [45]. Recording of these patient-reported outcome measures (PROMs) in the EHR is highly variable. New technologies, such as wearable devices and health-related apps are increasingly being transferred to EHRs [46, 47], often with limited or little attention to the quality, accuracy, and reliability of these data [8, 48].

Finally, some of the unstructured information in the EHR is based on the clinician’s discretion and might be subject to implicit bias [5]. Clinicians’ unintentional judgments and evaluations about patients’ attributes such as race or gender may affect what is documented or omitted [49]. Patients and their families have identified errors [50–52] and found offense in EHRs [34].

5.1.3 Peripheral Documents

Peripheral documents are mainly composed of imaging (x-rays, computerized tomography [CT] scans, magnetic resonance imaging [MRI] scans, ultrasounds, etc.) and non-numerical test results (electrocardiograms [ECGs], pathology results, etc.) as well as other portable document files or PDF files (scanned documents from sources outside the EHR). Extracting this information through NLP can provide a better understanding of the disease status of the patient than reliance on structured diagnostic codes [39]. However, these documents are not easily extractable [10].

5.2 Data Missingness

Structured data elements in EHRs may also be incompletely recorded [30, 53]. Structured data elements are likely to be most complete for health status and clinical information [10], but elements such as substance use and family history of illness, for example, are often missing; and absence of the entry cannot be assumed to represent absence of the behavior or condition [35]. Missingness can be at random, but can also arise due to the absence of data not intended to be collected (missing not at random) [7]. For example, some health information does not get recorded in EHR since it is

not relevant to some patients. This is known as “*informative missingness*,” when the absence of data conveys information about why it is missing [20]. Laboratory testing requested for some patients, for instance, suggests an underlying health condition and hence provides information about the health status of the patient. Not only do the actual values of the laboratory measurement carry information, but so do the number of measurements, the indications for the multiple measurements, and the dates of the measurements [23, 54].

Missing data in EHR can often lead to selection bias (by using complete data only) and misclassification bias [37]. Statistical techniques for missing data exist [30, 32, 37, 55, 56], but adjustments, such as imputation, cannot fully compensate for the bias in all circumstances [57–59].

Variables used to address confounding, such as sociodemographic status, health risk behaviors (diet, physical activity, substance use), and psychological stress are sometimes infrequently recorded [20, 60] and, when recorded, are usually of lower quality than data collected for research, which presents a challenge in accounting for confounders in EHR research. The FDA recommends linking EHR data to other data sources or attempting to collect additional data to capture information about important unmeasured or improperly measured confounders [7]. There are also several statistical techniques that attempt to address this problem, including inverse probability weighting [61], instrumental variable analysis [62–65], difference-in-differences [62, 66], interrupted time series [67], perturbation variable adjustment [63, 68], propensity scores calibration [63, 69–72], and sensitivity analysis [73].

6 Biases Arising at the Research Level

Biases can start from the very step of extracting the research database all the way to interpreting and reporting research findings.

7 Extraction and Acquisition of Research Data

Although Epic, Cerner, and Meditech are the most common EHR software in the US, hundreds of vendors develop EHR software, all of which have no common or uniform format [39, 47]. The structure of the EHRs is mainly vendor-specific [4] and differences in user interface can affect what data can be recorded and extracted from the database [17, 35]. Since the data extraction software is usually proprietary, the limitations and biases inherent in the software are nearly impossible to determine. Confidentiality agreements with EHR vendors make it difficult to evaluate the quality of the data extracted and the accuracy of the extraction tool [17, 35], and different extraction procedures may produce different results [74]. The challenges are further exacerbated by integrating different EHR software that lack interoperability or “*the ability of two or more products, technologies, or systems to exchange information and to use the information that has been exchanged without special effort on the part of the user*” [75] which also makes it difficult to track patients and identify duplication of patient records across multiple EHR systems. This is especially true because,

unlike the situation in some European countries, no single identification number is available to be used across the healthcare systems in the United States [9].

Further, raw EHR data can be transformed to research-grade variables through operational phenotyping algorithms [39, 76]. Two types of algorithms help define the study variables. The first involves the creation of a rule-based algorithm and sequential flow chart that identifies codes and clinical information in structured and unstructured fields. This method can be lengthy, might not use all the information available in EHR, and can be biased by the judgments of the team creating these algorithms [76]. The other type involves using machine learning algorithms [77], which has its own challenges and biases [4, 6, 39].

8 Research Design and Analysis

We list here several specific biases that can arise because of design and/or analysis issues in EHR research [78–80]. The list is not exhaustive but covers the most common biases affecting validity in EHR-derived research. Solutions to address these biases are suggested in the references cited for each one.

8.1 Berkson Bias

Berkson bias, or admission rate bias, is a type of selection bias resulting from the fact that patients with more than one condition are more likely to be hospitalized than patients with just one condition, creating a spurious correlation between diseases that are independent in the general population [81–83].

8.2 Informed Presence Bias

Analogous to Berkson's bias [84, 85], informed presence bias is a consequence of the nonrandom presence of patients and their details in the EHR system. People who have health challenges are more likely than others to seek care and have more medical encounters [24, 84]. When using data from these visits [86], a researcher unintentionally conditions on presence in the study sample [24, 84]. A prevalent but under-reported disease in the population will be more commonly documented in the sick because they are monitored more closely [87].

8.3 Prevalent-User Bias

Prevalent-user bias is a form of selection bias that is mainly present in drug effect studies. The bias arises when prevalent-users are compared to non-users especially when the treatment effects or the hazard of developing the outcomes vary with time [25, 88]. Prevalent-users are usually considered to be more tolerant of the treatment; i.e. they have “*survived*” early use [25, 88]. Therefore, if the risk of treatment-related outcomes is highest at the beginning of the treatment, the prevalent-user sample will consist of less susceptible patients and will consequently favor treatment [25, 80, 88].

8.4 Immortal Time Bias

Immortal time bias arises when a time interval exists between the assigned time of entry to the study and the time of exposure assignment, and tends to be found often among real-world cohort studies assessing treatment/drug effects [89]. This waiting time requires exposed participants to stay “immortal” and outcome-free until treatment assignment. Participants who experience the outcome before they have the chance to receive the treatment will be classified as unexposed [79, 89–91]. Consequently, the exposed group has a built-in survival advantage compared to the unexposed group and will appear to be protected, but the protection is artificial [89].

8.5 Lag-Time Bias

Lag-time bias is also a time-dependent bias, but one that relates to follow-up time after exposure assignment [88, 92]. The principal idea of the lag-time bias is that the risk of manifesting the outcome might not start immediately after the onset of exposure. Similarly, the risk might not end immediately after termination of the post-exposure observation interval.

8.6 Verification Bias

Verification bias occurs “when there is a difference in testing strategy between groups of individuals, leading to differing ways of verifying the disease of interest” [93]. It arises either when all patients receive an index test but only a proportion of them continue to receive the reference test for disease verification, or when patients are allocated to one of two reference tests based on the results of their index test [94].

In essence, verification bias occurs when the patients are not randomly selected to receive the reference test. In real-world clinical practice, factors like cost, invasiveness, individual patient susceptibility to risks as well as their preferences, and other healthcare system factors can all play a role in the nonrandom assignment of patients to subsequent diagnostic testing [95, 96].

8.7 Protopathic Bias

Protopathic bias occurs when a treatment/drug is prescribed to treat early signs and symptoms of a disease that has not yet been diagnosed [97–101]. It is the erroneous assumption that the drug caused the outcome when in fact the outcome gave rise to the treatment, a form of “reverse causation” [102, 103].

8.8 Confounding by Indication

Although not a bias per se [88, 104, 105], confounding by indication is the misinterpretation of an association between a drug/treatment and an outcome when the indication for selecting the drug/treatment contributes to the outcome [25, 106, 107]. The “indications” or reasons for treatment, such as the severity of the disease [88, 104–107], the frailty of the patient [88, 107], the physician’s preference for this drug

for this patient [104, 105], are considered confounders since they are associated with both the treatment and the outcome [105, 106].

9 Research Results and Interpretation

Since most EHR data are collected without *a-priori* research questions, their validity, relevance, and fitness for the specific research question need to be assessed [108]. Some analyses, however, go beyond this exploratory step, leading to modifications of pre-established study elements (i.e. inclusion/exclusion criteria, variable selection, variable definition, and analysis plan) that yield only favorable results [9, 109]. Up-front transparency about the research protocol, therefore, is instrumental in evaluating the quality and validity of the study. Using reporting guidelines like RECORD (Reporting of Studies Conducted using Observational Routinely Collected Health Data), MINORS (Methodological index for non-randomized studies), GRACE (The Good Research for Comparative Effectiveness), HARPER (HARmonized Protocol Template to Enhance Reproducibility), and STaRT-RWE (Structured Template for planning and Reporting on the implementation of Real World Evidence studies) are recommended for better transparency [2, 108–113]. Also, registration of RWE studies in publicly available databases prior to the execution of the study is likely to maximize transparency [109].

The Newcastle-Ottawa Quality Assessment Scale [114] provides a composite score of assessments of the representativeness of non-randomized studies, the quality of the exposure and outcome, avoidance of biases, and adjustments in the analyses. As a composite score it has the potential to avoid information overload [115], and to intensify the signal of interest [116]. With these advantages comes the potential for weighing components in less than-desirable ways [117, 118]. Consequently, we advise paying attention to the components, as well as the score.

10 Challenges of Pooling Multi-Institutional EHRs

While pooling EHR databases is recommended for validating of study variables [24, 25], minimizing information bias due to EHR discontinuity [25], and increasing power [1], doing so adds another level of complexity to the above-mentioned challenges encountered within an individual EHR system. Integrating multi-institutional EHRs data requires careful evaluation of the heterogeneity in medical practices, reimbursement systems, organizational policies, and demographic characteristics of the catchment areas that can impact the type and quality of data captured [5, 9, 35]. For example, the patient profiles in academic healthcare systems, suburban practices, and federally-qualified health centers are likely to differ [5]. Such heterogeneity can impact the type and quality of data captured, which can, in turn, complicate the integration process for these EHR databases.

Pooling multiple EHR systems is probably best achieved by a multidisciplinary team of clinicians, scientists, informaticians, ML experts, ethical experts who know about the practices of their healthcare system and how data are captured and recorded

at their own institutions [35]. Adjusting for clustering or center effects can reduce bias, but is unlikely to eliminate it [119]. Finally, even when data precision, completeness, interoperability, and harmonization are addressed, security and patient privacy remain a concern [44, 78].

In lieu of the challenges of pooling multiple EHR databases, several initiatives have created a network of RWD that aim to collect and map information to a common data model (CDM) that has a consistent format and content [9]. It aims to standardize data collection across several EHR systems and facilitate system interoperability and data sharing [39]. Some of the main initiatives are the Observational Health Data Sciences and Informatics (OHDSI) [120], the National Patient-Centered Clinical Research Network (PCORnet) [121], and, most recently the National COVID Cohort Collaborative (N3C) consortium that emerged as a response to the COVID19 pandemic [122]. Although these initiatives are promising for establishing standardized RWD, they are still under development [4].

11 The Future

More recently, the Office of the National Coordinator for Health, in the US Department of Health and Human Services, released the Trusted Exchange Framework and Common Agreement (TEFCA) which enables the creation of nationwide data-sharing networks, known as Qualified Health Information Networks (QHIN). These QHINs are expected to connect to one another to support a “*network of networks*,” resulting in national health information exchange. The TEFCA also aims to create a common set of practices to promote homogeneity of data collected in EHRs [123, 124]. Among these is federated learning [125, 126], which enables multiple contributors “*to build a common, robust machine learning model without sharing data, thus addressing critical issues such as data privacy, data security, data access rights and access to heterogeneous data*” [127]. With the implementation of the TEFCA, we can probably expect a prominent increase in the use of RWE for research and post-marketing surveillance.

12 Conclusion

Now, more than ever, EHRs are being perceived as unique sources of data for clinical research, providing unprecedentedly large volumes of real-time data from real-world settings. Solely having access to big data, however, does not minimize or eliminate bias. Because having a large sample size increases statistical precision, flawed big data increases the chances of significantly biased inferences; this is a phenomenon known as the “*big data paradox*” [19, 20]. Large datasets, therefore, do not necessarily lead to quality research and valid RWE.

In summary, EHRs might not provide a complete reflection of the patient and his/her health status. Instead, they are a reflection of the utilization of healthcare services and EHR recording processes [76]. Limited data quality and the plethora of biases in EHR data prompt us to conclude with words of caution. As appealing as EHR

data might appear, we recommend investigators carefully design their study with the above-mentioned challenges in mind. Even after what might be considered extreme efforts to maximize data quality and minimize bias, humility is encouraged.

Acknowledgements We would like to thank Dr. Maitreyi Mazumdar from Harvard Medical School for supporting and revising this manuscript.

Authors' Contributions BAS: Made substantial contribution to the conception and design of the paper; performed literature review and prepared all the drafts of the paper. AL: Made substantial contribution to the conception of the paper and design of the paper; performed literature review; provided technical support and advice; and contributed to the write up of the first draft of the paper; critically revised the final paper and all its drafts for important intellectual content. TL: Made substantial contribution to the conception of the paper and design of the paper; provided technical support and advice; critically revised the final paper and all its drafts for important intellectual content. NP: Made substantial contribution to the conception of the paper and design of the paper; provided technical support and advice; critically revised the final paper and all its drafts for important intellectual content. BZ: Made substantial contribution to the conception of the paper and design of the paper; provided technical and statistical support and advice; critically revised the final paper and all its drafts for important intellectual content.

Funding No funding was available for this manuscript. The author BAS prepared this manuscript while completing her postdoctoral training in the Department of Epidemiology and Biostatistics supported by the Office of Research in Women's Health, National Institute of Health under Award Number 3UH3OD023285-06S1, linked to the parent award UH3OD023285.

Declarations

Competing Interests None of the authors have any conflict of interest regarding the material discussed in the manuscript.

References

1. Sandhu E, Weinstein S, McKethan A, Jain SH (2012) Secondary uses of electronic health record data: benefits and barriers. *Jt Comm J Qual Patient Saf* 38(1):34–40. [https://doi.org/10.1016/s1553-7250\(12\)38005-7](https://doi.org/10.1016/s1553-7250(12)38005-7)
2. Liu M, Qi Y, Wang W, Sun X (2022) Toward a better understanding about real-world evidence. *Eur J Hosp Pharm* 29(1):8–11. <https://doi.org/10.1136/ejhpharm-2021-003081>
3. Concato J, Corrigán-Curay J (2022) Real-world evidence - where are we now? *N Engl J Med* 386(18):1680–1682. <https://doi.org/10.1056/NEJMp2200089>
4. Holmes JH, Beinlich J, Boland MR, Bowles KH, Chen Y, Cook TS, Demiris G, Draugelis M, Fluharty L, Gabriel PE et al (2021) Why is the Electronic Health Record so challenging for Research and Clinical Care? *Methods Inf Med* 60(1–02):32–48. <https://doi.org/10.1055/s-0041-1731784>
5. Gianfrancesco MA, Goldstein ND (2021) A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol* 21(1):234. <https://doi.org/10.1186/s12874-021-01416-5>
6. Knevel R, Liao KP (2023) From real-world electronic health record data to real-world results using artificial intelligence. *Ann Rheum Dis* 82(3):306–311. <https://doi.org/10.1136/ard-2022-222626>
7. Food US, and Drug Administration (FDA) (2021). Real-World Data: Assessing Electronic Health Records Medical Claims Data To Support Regulatory Decision-Making for Drug Biological Products. <https://www.fda.gov/media/152503/download>. Accessed June 2023
8. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Hunter NL, LaVange L, Marinac-Dabic D, Marks PW, Robb MA et al (2016) Real-world evidence - what is it and what can it tell us? *N Engl J Med* 375(23):2293–2297. <https://doi.org/10.1056/NEJMs1609216>

9. U.S. Food and Drug Administration (FDA). Framework for FDA's Real-World Evidence Program. <https://www.fda.gov/media/120060/download>. (2018) Accessed June 2023
10. Duke-Margolis Center for Health Policy (2019) Determining Real-World Data's Fitness for Use and the Role of Reliability. https://healthpolicy.duke.edu/sites/default/files/2019-11/rwd_reliability.pdf. Accessed June 2023
11. Singhal P, Tan ALM, Drivas TG, Johnson KB, Ritchie MD (2023) Beaulieu-Jones: opportunities and challenges for biomarker discovery using electronic health record data. *Trends Mol Med* 29(9):765–776. <https://doi.org/10.1016/j.molmed.2023.06.006>
12. Pasternak AL, Ward K, Irwin M, Okerberg C, Hayes D, Fritsche L, Zoellner S, Virzi J, Choe HM, Ellingrod V (2023) Identifying the prevalence of clinically actionable drug-gene interactions in a health system biorepository to guide pharmacogenetics implementation services. *Clin Transl Sci* 16(2):292–304. <https://doi.org/10.1111/cts.13449>
13. Zhao Y, Tsubota T (2023) The current status of secondary use of claims, Electronic Medical Records, and Electronic Health Records in Epidemiology in Japan: Narrative Literature Review. *JMIR Med Inform* 11. <https://doi.org/10.2196/39876>
14. Iott BE, Adler-Milstein J, Gottlieb LM, Pantell MS (2023) Characterizing the relative frequency of clinician engagement with structured social determinants of health data. *J Am Med Inform Assoc* 30(3):503–510. <https://doi.org/10.1093/jamia/ocac251>
15. Dixit RA, Boxley CL, Samuel S, Mohan V, Ratwani RM, Gold JA (2023) Electronic Health Record Use issues and Diagnostic Error: a scoping review and Framework. *J Patient Saf* 19(1):e25–e30. <https://doi.org/10.1097/pts.0000000000001081>
16. Modi S, Feldman SS (2022) The Value of Electronic Health Records since the Health Information Technology for Economic and Clinical Health Act: systematic review. *JMIR Med Inform* 10(9):e37283. <https://doi.org/10.2196/37283>
17. Verheij RA, Curcin V, Delaney BC, McGilchrist MM (2018) Possible sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *J Med Internet Res* 20(5):e185. <https://doi.org/10.2196/jmir.9134>
18. Last JM (1983) *A Dictionary of Epidemiology*. Oxford University Press, United Kingdom
19. Beesley LJ, Mukherjee B (2022) Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics* 78(1):214–226. <https://doi.org/10.1111/biom.13400>
20. Bots SH, Groenwold RHH, Dekkers OM (2022) Using electronic health record data for clinical research: a quick guide. *Eur J Endocrinol* 186(4):E1–e6. <https://doi.org/10.1530/eje-21-1088>
21. Romo ML, Chan PY, Lurie-Moroni E, Perlman SE, Newton-Dame R, Thorpe LE, McVeigh KH (2016) Characterizing adults receiving Primary Medical Care in New York City: Implications for Using Electronic Health Records for Chronic Disease Surveillance. *Prev Chronic Dis* 13. <https://doi.org/10.5888/pcd13.150500>
22. Phelan M, Bhavsar NA, Goldstein BA (2017) EGEMS (Wash DC) 5(1):22. <https://doi.org/10.5334/egems.243>. Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference
23. Agniet D, Kohane IS, Weber GM (2018) Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* k1479. 36110.1136/bmj.k1479
24. Bower JK, Patel S, Rudy JE, Felix AS (2017) Addressing Bias in Electronic Health Record-based surveillance of Cardiovascular Disease Risk: finding the Signal through the noise. *Curr Epidemiol Rep* 4(4):346–352. <https://doi.org/10.1007/s40471-017-0130-z>
25. Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L (2018) Promises and pitfalls of electronic health record analysis. *Diabetologia* 61(6):1241–1248. <https://doi.org/10.1007/s00125-017-4518-6>
26. Williams BA (2021) Constructing epidemiologic cohorts from Electronic Health Record Data. *Int J Environ Res Public Health* 18(24). <https://doi.org/10.3390/ijerph182413193>
27. Beesley LJ, Mukherjee B (2020) : Bias reduction and inference for electronic health record data under selection and phenotype misclassification: three case studies. medRxiv <https://doi.org/10.1101/2020.12.21.20248644>
28. Goldstein ND, Kahal D, Testa K, Gracely EJ, Burstyn I (2022) : Data Quality in Electronic Health Record Research: an Approach for Validation and Quantitative Bias Analysis for Imperfectly Ascertained Health outcomes Via Diagnostic codes. *Harv Data Sci Rev* 4(2)

29. 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* (2016), 37, p. 61–81. <https://doi.org/10.1146/annurev-publhealth-032315-021353>
30. Peskoe SB, Arterburn D, Coleman KJ, Herrinton LJ, Daniels MJ, Haneuse S (2021) Adjusting for selection bias due to missing data in electronic health records-based research. *Stat Methods Med Res* 30(10):2221–2238. <https://doi.org/10.1177/09622802211027601>
31. Jin Y, Schneeweiss S, Merola D, Lin KJ (2022) Impact of longitudinal data-completeness of electronic health record data on risk score misclassification. *J Am Med Inform Assoc* 29(7):1225–1232. <https://doi.org/10.1093/jamia/ocac043>
32. Haneuse S, Daniels M (2016) : A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why? *EGEMS* (Wash DC) 4(1), p. 1203. <https://doi.org/10.13063/2327-9214.1203>
33. Congressional Research Service (CRS) (2016) The 21st Century Cures Act (Division A of P.L. 114–255). <https://sgp.fas.org/crs/misc/R44720.pdf>. Accessed June 2023
34. Fernández L, Fossa A, Dong Z, Delbanco T, Elmore J, Fitzgerald P, Harcourt K, Perez J, Walker J, DesRoches C (2021) Words Matter: what do patients find judgmental or Offensive in Outpatient notes? *J Gen Intern Med* 36(9):2571–2578. <https://doi.org/10.1007/s11606-020-06432-7>
35. Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, Brat GA, Cannataro M, Cimino JJ, García-Barrio N et al (2021) What every reader should know about studies using Electronic Health Record Data but May be afraid to ask. *J Med Internet Res* 23(3):e22219. <https://doi.org/10.2196/22219>
36. Beesley LJ, Fritsche LG, Mukherjee B (2020) An analytic framework for exploring sampling and observation process biases in genome and phenome-wide association studies using electronic health records. *Stat Med* 39(14):1965–1979. <https://doi.org/10.1002/sim.8524>
37. Khurshid S, Reeder C, Harrington LX, Singh P, Sarma G, Friedman SF, Di Achille P, Diamant N, Cunningham JW, Turner AC et al (2022) Cohort design and natural language processing to reduce bias in electronic health records research. *NPJ Digit Med* 5(1):47. <https://doi.org/10.1038/s41746-022-00590-0>
38. Huang J, Duan R, Hubbard RA, Wu Y, Moore JH, Xu H, Chen Y (2018) PIE: a prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *J Am Med Inform Assoc* 25(3):345–352. <https://doi.org/10.1093/jamia/ocx137>
39. Pendergrass SA, Crawford DC (2019) Using Electronic Health Records To Generate Phenotypes for Research. *Curr Protoc Hum Genet* 100(1):e80. <https://doi.org/10.1002/cphg.80>
40. Agency for Healthcare Research and Quality (AHRQ) (2019) Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User’s Guide, 3rd Edition, Addendum 2. https://www.ncbi.nlm.nih.gov/books/NBK551879/pdf/Bookshelf_NBK551879.pdf. Accessed June 2023
41. Muntner P, Einhorn PT, Cushman WC, Whelton PK, Bello NA, Drawz PE, Green BB, Jones DW, Juraschek SP, Margolis KL et al (2019) Blood pressure Assessment in adults in clinical practice and clinic-based research: JACC Scientific Expert Panel. *J Am Coll Cardiol* 73(3):317–335. <https://doi.org/10.1016/j.jacc.2018.10.069>
42. Kim HS, Kim JH (2019) *J Korean Med Sci* 34(4):e28. <https://doi.org/10.3346/jkms.2019.34.e28>. Proceed with Caution When Using Real World Data and Real World Evidence
43. van der Bij S, Khan N, Ten Veen P, de Bakker DH, Verheij RA (2017) Improving the quality of EHR recording in primary care: a data quality feedback tool. *J Am Med Inform Assoc* 24(1):81–87. <https://doi.org/10.1093/jamia/ocw054>
44. Cowie MR, Blomster JJ, Curtis LH, Duclaux S, Ford I, Fritz F, Goldman S, Janmohamed S, Kreuzer J, Leenay M et al (2017) Electronic health records to facilitate clinical research. *Clin Res Cardiol* 106(1):1–9. <https://doi.org/10.1007/s00392-016-1025-6>
45. Brinkmann BH, Karoly PJ, Nurse ES, Dumanis SB, Nasserli M, Viana PF, Schulze-Bonhage A, Freestone DR, Worrell G, Richardson MP et al (2021) Seizure diaries and forecasting with wearables: Epilepsy Monitoring outside the clinic. *Front Neurol* 12:690404. <https://doi.org/10.3389/fneur.2021.690404>
46. Shaw R, Stroom M, Fiander C, McMillan K (2020) Selecting Mobile Health Technologies for Electronic Health Record Integration: Case Study. *J Med Internet Res* 22(10):e23314. <https://doi.org/10.2196/23314>

47. Dinh-Le C, Chuang R, Chokshi S, Mann D (2019) Wearable Health Technology and Electronic Health Record Integration: scoping review and future directions. *JMIR Mhealth Uhealth* 7(9):e12861. <https://doi.org/10.2196/12861>
48. Collins T, Woolley SI, Oniani S, Pandyan A (2021) Quantifying missingness in Wearable Heart Rate recordings. *Stud Health Technol Inform* 281:1077–1078. <https://doi.org/10.3233/SHTI210352>
49. Sun M, Oliwa T, Peek ME, Tung EL (2022) Negative patient descriptors: documenting racial Bias in the Electronic Health Record. *Health Aff (Millwood)* 41(2):203–211. <https://doi.org/10.1377/hlthaff.2021.01423>
50. Bourgeois FC, Fossa A, Gerard M, Davis ME, Taylor YJ, Connor CD, Vaden T, McWilliams A, Spencer MD, Folcarelli P et al (2019) A patient and family reporting system for perceived ambulatory note mistakes: experience at 3 U.S. healthcare centers. *J Am Med Inform Assoc* 26(12):1566–1573. <https://doi.org/10.1093/jamia/ocz142>
51. Lam BD, Bourgeois F, Dong ZJ, Bell SK (2021) Speaking up about patient-perceived serious visit note errors: patient and family experiences and recommendations. *J Am Med Inform Assoc* 28(4):685–694. <https://doi.org/10.1093/jamia/ocaa293>
52. Lear R, Freise L, Kybert M, Darzi A, Neves AL, Mayer EK (2022) Patients' willingness and ability to identify and respond to errors in their Personal Health records: mixed methods analysis of cross-sectional Survey Data. *J Med Internet Res* 24(7):e37226. <https://doi.org/10.2196/37226>
53. Haneuse S, Bogart A, Jazic I, Westbrook EO, Boudreau D, Theis MK, Simon GE, Arterburn D (2016) Learning About Missing Data Mechanisms in Electronic Health Records-based Research: a Survey-based Approach. *Epidemiology* 27(1):82–90. <https://doi.org/10.1097/ede.0000000000000393>
54. Little RJA, Rubin DB (2019) *Statistical analysis with Missing Data*. John Wiley & Sons, New York, NY
55. Groenwold RHH (2020) Informative missingness in electronic health record systems: the curse of knowing. *Diagn Progn Res* 8. 410.1186/s41512-020-00077-0
56. Haneuse S, Arterburn D, Daniels MJ (2021) Assessing Missing Data assumptions in EHR-Based studies: a Complex and Underappreciated Task. *JAMA Netw Open* 4(2):e210184. <https://doi.org/10.1001/jamanetworkopen.2021.0184>
57. Ford E, Rooney P, Hurley P, Oliver S, Bremner S, Cassell J (2020) Can the use of Bayesian Analysis Methods Correct for Incompleteness in Electronic Health Records Diagnosis Data? Development of a Novel Method using simulated and real-life Clinical Data. *Front Public Health* 8. <https://doi.org/10.3389/fpubh.2020.00054>
58. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR (2000) Bayesian methods in health technology assessment: a review. *Health Technol Assess* 4(38):1–130
59. Li J, Yan XS, Chaudhary D, Avula V, Mudiganti S, Husby H, Shahjouei S, Afshar A, Stewart WF, Yeasin M et al (2021) Imputation of missing values for electronic health record laboratory data. *NPJ Digit Med* 4(1):147. <https://doi.org/10.1038/s41746-021-00518-0>
60. Cook LA, Sachs J, Weiskopf NG (2021) The quality of social determinants data in the electronic health record: a systematic review. *J Am Med Inform Assoc* 29(1):187–196. <https://doi.org/10.1093/jamia/ocab199>
61. Sayon-Orea C, Moreno-Iribas C, Delfrade J, Sanchez-Echenique M, Amiano P, Ardanaz E, Gorricho J, Basterra G, Nuin M, Guevara M (2020) Inverse-probability weighting and multiple imputation for evaluating selection bias in the estimation of childhood obesity prevalence using data from electronic health records. *BMC Med Inform Decis Mak* 20(1):9. <https://doi.org/10.1186/s12911-020-1020-8>
62. Streeter AJ, Lin NX, Crathorne L, Haasova M, Hyde C, Melzer D (2017) and W.E. Henley: Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *J Clin Epidemiol* 87, p. 23–34. <https://doi.org/10.1016/j.jclinepi.2017.04.022>
63. Uddin MJ, Groenwold RH, Ali MS, de Boer A, Roes KC, Chowdhury MA, Klungel OH Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *Int J Clin Pharm* (2016), 38(3), p. 714–723. <https://doi.org/10.1007/s11096-016-0299-0>
64. Richardson DB, Tchetgen Tchetgen EJ (2021) Bespoke instruments: a new tool for addressing unmeasured confounders. *Am J Epidemiol*. <https://doi.org/10.1093/aje/kwab288>
65. Krishnamoorthy V, McLean D, Ohnuma T, Harris SK, Wong DJN, Wilson M, Moonesinghe R, Raghunathan K (2020) Causal inference in perioperative medicine observational research: part 2, advanced methods. *Br J Anaesth* 125(3):398–405. <https://doi.org/10.1016/j.bja.2020.03.032>
66. Craig P, Katikireddi SV, Leyland A, Popham F (2017) Natural experiments: an overview of methods, approaches, and contributions to Public Health Intervention Research. *Annu Rev Public Health* 38:39–56. <https://doi.org/10.1146/annurev-publhealth-031816-044327>

67. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ* (2015), 350, p. h2750. <https://doi.org/10.1136/bmj.h2750>
68. Lee WC (2014) Detecting and correcting the bias of unmeasured factors using perturbation analysis: a data-mining approach. *BMC Med Res Methodol* 14:18. <https://doi.org/10.1186/1471-2288-14-18>
69. Balalian AA, Daniel S, Simonyan H, Khachadourian V (2022) Comparison of conditional and marginal models in assessing a child Nutrition intervention in Armenia. *Matern Child Health*. <https://doi.org/10.1007/s10995-021-03308-y>
70. Fujiwara Y, Fukuda S, Tsujie M, Kitani K, Yukawa M, Inoue M, Watanabe Y, Higashida M, Kubota H, Okada T et al (2019) Clinical significance of preoperative chemoradiotherapy for advanced Esophageal cancer, evaluated by propensity score matching and weighting of inverse probability of treatment. *Mol Clin Oncol* 10(6):575–582. <https://doi.org/10.3892/mco.2019.1843>
71. Allan V, Ramagopalan SV, Mardekian J, Jenkins A, Li X, Pan X, Luo X (2020) Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants. *J Comp Eff Res* 9(9):603–614. <https://doi.org/10.2217/ceer-2020-0013>
72. Austin PC, Stuart EA (2017) The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Stat Methods Med Res* 26(4):1654–1670. <https://doi.org/10.1177/0962280215584401>
73. Schneeweiss S (2006) Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf* 15(5):291–303. <https://doi.org/10.1002/pds.1200>
74. Liaw ST, Taggart J, Yu H, de Lusignan S (2013) Data extraction from electronic health records - existing tools may be unreliable and potentially unsafe. *Aust Fam Physician* 42(11):820–823
75. U.S. Food and Drug Administration (FDA) (2018) Use of Electronic Health Record Data in Clinical Investigations Guidance for Industry. <https://www.fda.gov/media/97567/download>. Accessed June 2023
76. Hripcsak G, Albers DJ (2013) Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 20(1):117–121. <https://doi.org/10.1136/amiajnl-2012-001145>
77. Rowe M (2019) An introduction to machine learning for clinicians. *Acad Med* 94(10):1433–1436. <https://doi.org/10.1097/acm.0000000000002792>
78. Nair S, Hsu D, Celi LA (2016) Challenges and opportunities in Secondary Analyses of Electronic Health Record Data, in *Secondary Analysis of Electronic Health Records*. Springer, Cham (CH), pp 17–26
79. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* (2016), 79, p. 70–75. <https://doi.org/10.1016/j.jclinepi.2016.04.014>
80. Hernán MA, Robins JM (2016) Using Big Data to emulate a target Trial when a Randomized Trial is not available. *Am J Epidemiol* 183(8):758–764. <https://doi.org/10.1093/aje/kwv254>
81. Snoep JD, Morabia A, Hernández-Díaz S, Hernán MA, Vandenbroucke JP (2014) Commentary: a structural approach to Berkson’s fallacy and a guide to a history of opinions about it. *Int J Epidemiol* 43(2):515–521. <https://doi.org/10.1093/ije/dyu026>
82. Berkson J (1946) Limitations of the application of fourfold table analysis to hospital data. *Biometrics* 2(3):47–53
83. Sackett DL (1979) Bias in analytic research. *J Chronic Dis* 32(1–2):51–63. [https://doi.org/10.1016/0021-9681\(79\)90012-2](https://doi.org/10.1016/0021-9681(79)90012-2)
84. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ (2016) Controlling for Informed Presence Bias due to the Number of Health Encounters in an Electronic Health Record. *Am J Epidemiol* 184(11):847–855. <https://doi.org/10.1093/aje/kww112>
85. Goldstein BA, Phelan M, Pagidipati NJ, Peskoe SB (2019) How and when informative visit processes can bias inference when using electronic health records data for clinical research. *J Am Med Inform Assoc* 26(12):1609–1617. <https://doi.org/10.1093/jamia/ocz148>
86. Harton J, Mitra N, Hubbard RA (2022) Informative presence bias in analyses of electronic health records-derived data: a cautionary note. *J Am Med Inform Assoc* 29(7):1191–1199. <https://doi.org/10.1093/jamia/ocac050>

87. McGee G, Haneuse S, Coull BA, Weisskopf MG, Rotem RS (2022) On the Nature of Informative Presence Bias in Analyses of Electronic Health Records. *Epidemiology* 33(1):105–113. <https://doi.org/10.1097/ede.0000000000001432>
88. Gokhale M, Stürmer T, Buse JB (2020) Real-world evidence: the devil is in the detail. *Diabetologia* 63(9):1694–1705. <https://doi.org/10.1007/s00125-020-05217-1>
89. Suissa S (2008) Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol* 167(4):492–499. <https://doi.org/10.1093/aje/kwm324>
90. Tyrer F, Bhaskaran K, Rutherford MJ (2022) Immortal time bias for life-long conditions in retrospective observational studies using electronic health records. *BMC Med Res Methodol* 22(1):86. <https://doi.org/10.1186/s12874-022-01581-1>
91. Lévesque LE, Hanley JA, Kezouh A, Suissa S (2010) Problem of immortal time bias in cohort studies: example using statins for preventing progression of Diabetes. *BMJ* 340:b5087. <https://doi.org/10.1136/bmj.b5087>
92. Iudici M, Porcher R, Riveros C, Ravaud P (2019) Time-dependent biases in observational studies of comparative effectiveness research in rheumatology. A methodological review. *Ann Rheum Dis* 78(4):562–569. <https://doi.org/10.1136/annrheumdis-2018-214544>
93. CoB C Catalogue of Bias. Oxford: England, UK: University of Oxford
94. O'Sullivan JW, Banerjee A, Heneghan C, Pluddemann A (2018) Verification bias. *BMJ Evid Based Med* 23(2):54–55. <https://doi.org/10.1136/bmjebm-2018-110919>
95. de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Brophy J, Joseph L, Bossuyt PM, Moons KG (2012) Adjusting for partial verification or workup bias in meta-analyses of diagnostic accuracy studies. *Am J Epidemiol* 175(8):847–853. <https://doi.org/10.1093/aje/kwr383>
96. Brown CA, Londhe AA, He F, Cheng A, Ma J, Zhang J, Brooks CG, Sprafka JM, Roehl KA, Carlson KB et al (2022) Development and Validation of algorithms to identify COVID-19 patients using a US Electronic Health Records Database: a retrospective cohort study. *Clin Epidemiol* 14:699–709. <https://doi.org/10.2147/clep.S355086>
97. Horwitz RI, Feinstein AR (1980) The problem of protopathic bias in case-control studies. *Am J Med* 68(2):255–258. [https://doi.org/10.1016/0002-9343\(80\)90363-0](https://doi.org/10.1016/0002-9343(80)90363-0)
98. Singh A, Hussain S, Akkala S, Klugarová J, Pokorná A, Klugar M, Walters EH, Hopper I, Campbell JA, Taylor B et al (2022) Beta-adrenergic drugs and risk of Parkinson's disease: A systematic review and meta-analysis. *Ageing Res Rev* 80. <https://doi.org/10.1016/j.arr.2022.101670>
99. Tamim H, Monfared AA, LeLorier J (2007) Application of lag-time into exposure definitions to control for protopathic bias. *Pharmacoepidemiol Drug Saf* 16(3):250–258. <https://doi.org/10.1002/pds.1360>
100. Faillie JL (2015) Indication bias or protopathic bias? *Br J Clin Pharmacol* 80(4):779–780. <https://doi.org/10.1111/bcp.12705>
101. Prada-Ramallal G, Takkouche B, Figueiras A (2019) Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. *BMC Med Res Methodol* 19(1):53. <https://doi.org/10.1186/s12874-019-0695-y>
102. Murk W, Risnes KR, Bracken MB (2011) Prenatal or early-life exposure to antibiotics and risk of childhood Asthma: a systematic review. *Pediatrics* 127(6):1125–1138. <https://doi.org/10.1542/peds.2010-2092>
103. Lo CH, Ni P, Yan Y, Ma W, Joshi AD, Nguyen LH, Mehta RS, Lochhead P, Song M, Curhan GC et al (2022) : Association of Proton Pump Inhibitor Use With All-Cause and Cause-Specific Mortality. *Gastroenterology* 163(4), p. 852–861.e2. <https://doi.org/10.1053/j.gastro.2022.06.067>
104. Walker AM (1996) Confounding by indication. *Epidemiology* 7(4):335–336
105. Salas M, Hofman A, Stricker BH (1999) Confounding by indication: an example of variation in the use of epidemiologic terminology. *Am J Epidemiol* 149(11):981–983. <https://doi.org/10.1093/oxfordjournals.aje.a009758>
106. Kyriacou DN, Lewis RJ (2016) Confounding by indication in Clinical Research. *JAMA* 316(17):1818–1819. <https://doi.org/10.1001/jama.2016.16435>
107. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I (2013) Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 347:f6409. <https://doi.org/10.1136/bmj.f6409>
108. Wang SV, Schneeweiss S (2022) Assessing and interpreting real-world evidence studies: introductory points for New Reviewers. *Clin Pharmacol Ther* 111(1):145–149. <https://doi.org/10.1002/cpt.2398>

109. Orsini LS, Monz B, Mullins CD, Van Brunt D, Daniel G, Eichler HG, Graff J, Guerino J, Berger M, Lederer NM et al (2020) Improving transparency to build trust in real-world secondary data studies for hypothesis testing—Why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Pharmacoepidemiol Drug Saf* 29(11):1504–1513. <https://doi.org/10.1002/pds.5079>
110. Dreyer NA, Bryant A, Velentgas P (2016) The GRACE Checklist: a validated Assessment Tool for High Quality Observational studies of comparative effectiveness. *J Manag Care Spec Pharm* 22(10):1107–1113. <https://doi.org/10.18553/jmcp.2016.22.10.1107>
111. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, Sorensen HT, von Elm E, Langan SM (2015) The REporting of studies conducted using Observational routinely-collected health data (RECORD) statement. *PLoS Med* 12(10):e1001885. <https://doi.org/10.1371/journal.pmed.1001885>
112. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J (2003) Methodological index for non-randomized studies (minors): development and validation of a new instrument. *ANZ J Surg* 73(9):712–716. <https://doi.org/10.1046/j.1445-2197.2003.02748.x>
113. Wang SV, Pottegård A, Crown W, Arlett P, Ashcroft DM, Benchimol EI, Berger ML, Crane G, Goetsch W, Hua W et al (2023) HARMonized Protocol Template to enhance reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: a good practices report of a joint ISPE/ISPOR task force. *Pharmacoepidemiol Drug Saf* 32(1):44–55. <https://doi.org/10.1002/pds.5507>
114. Wells GA, Shea B, O’Connell D et al (2021) The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed June 2023
115. Friebel R, Steventon A (2019) Composite measures of healthcare quality: sensible in theory, problematic in practice. *BMJ Qual Saf* 28(2):85–88. <https://doi.org/10.1136/bmjqs-2018-008280>
116. Austin PC, Ceyisakar IE, Steyerberg EW, Lingsma HF (2019) Marang-Van De Mheen: ranking hospital performance based on individual indicators: can we increase reliability by creating composite indicators? *BMC Med Res Methodol* 19(1):131. <https://doi.org/10.1186/s12874-019-0769-x>
117. Greco S, Ishizaka A, Tasiou M, Torrisi G (2019) On the Methodological Framework of Composite indices: a review of the issues of weighting, aggregation, and Robustness. *Soc Indic Res* 141(1):61–94. <https://doi.org/10.1007/s11205-017-1832-9>
118. Kara P, Valentin JB, Mainz J, Johnsen SP (2022) Composite measures of quality of health care: evidence mapping of methodology and reporting. *PLoS ONE* 17(5):e0268320. <https://doi.org/10.1371/journal.pone.0268320>
119. Localio AR, Berlin JA, Ten TR, Have, Kimmel SE (2001) Adjustments for center in multicenter studies: an overview. *Ann Intern Med* 135(2):112–123. <https://doi.org/10.7326/0003-4819-135-2-200107170-00012>
120. The Observational Health Data Sciences and Informatics (2023) <https://www.ohdsi.org/>. Accessed
121. National Patient-Centered Clinical Research Network (2023) <https://pcornet.org/>. Accessed
122. N3C (2023) <https://covid.cd2h.org/>. Accessed
123. The Office of the National Coordinator for Health Information Technology (ONC) (2023) Trusted Exchange Framework and Common Agreement (TEFCA). <https://www.healthit.gov/topic/interoperability/policy/trusted-exchange-framework-and-common-agreement-tefca>. Accessed
124. Mandel JC, Pollak JP, Mandl KD (2022) The patient role in a Federal National-Scale Health Information Exchange. *J Med Internet Res* 24(11):e41750. <https://doi.org/10.2196/41750>
125. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, Milchenko M, Xu W, Marcus D, Colen RR et al (2020) Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep* 10(1):12598. <https://doi.org/10.1038/s41598-020-69250-1>
126. Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, Foley P, Gruzdev A, Karkada D, Davatzikos C et al (2022) Federated learning enables big data for rare cancer boundary detection. *Nat Commun* 13(1):7346. <https://doi.org/10.1038/s41467-022-33407-5>
127. Wikipedia (2023) Federated Learning. https://en.wikipedia.org/wiki/Federated_learning Accessed

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted

manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Ban Al-Sahab¹ · Alan Leviton^{2,3} · Tobias Loddenkemper^{2,3} · Nigel Paneth^{4,5} · Bo Zhang^{3,6,7}

✉ Ban Al-Sahab
alsahab1@msu.edu

¹ Department of Family Medicine, College of Human Medicine, Michigan State University, B100 Clinical Center, 788 Service Road, East Lansing, MI, USA

² Department of Neurology, Harvard Medical School, Boston, MA, USA

³ Department of Neurology, Boston Children’s Hospital, Boston, MA, USA

⁴ Department of Epidemiology and Biostatistics, College of Human Medicine, Michigan State University, East Lansing, MI, USA

⁵ Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, East Lansing, MI, USA

⁶ Biostatistics and Research Design, Institutional Centers of Clinical and Translational Research, Boston Children’s Hospital, Boston, MA, USA

⁷ Harvard Medical School, Boston, MA, USA