Gut and Liver
https://doi.org/10.5009/gnl230272
pISSN 1976-2283 eISSN 2005-1212

Review Article

Challenges in and Opportunities for Electronic Health Record-Based Data Analysis and Interpretation

Michelle Kang Kim[1], Carol Rouphael[1], John McMichael[2], Nicole Welch[1,3], Srinivasan Dasarathy[1,3]

[1]Department of Gastroenterology, Hepatology, and Nutrition, Digestive Disease and Surgery Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA; [2]Department of Surgery, Digestive Disease and Surgery Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA; [3]Department of Inflammation and Immunity, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA

Article Info
Received July 14, 2023
Accepted August 15, 2023
Published online October 31, 2023

Corresponding Author
Michelle Kang Kim
ORCID https://orcid.org/0000-0001-5285-8218
E-mail kimm13@ccf.org

Electronic health records (EHRs) have been increasingly adopted in clinical practices across the United States, providing a primary source of data for clinical research, particularly observational cohort studies. EHRs are a high-yield, low-maintenance source of longitudinal real-world data for large patient populations and provide a wealth of information and clinical contexts that are useful for clinical research and translation into practice. Despite these strengths, it is important to recognize the multiple limitations and challenges related to the use of EHR data in clinical research. Missing data are a major source of error and biases and can affect the representativeness of the cohort of interest, as well as the accuracy of the outcomes and exposures. Here, we aim to provide a critical understanding of the types of data available in EHRs and describe the impact of data heterogeneity, quality, and generalizability, which should be evaluated prior to and during the analysis of EHR data. We also identify challenges pertaining to data quality, including errors and biases, and examine potential sources of such biases and errors. Finally, we discuss approaches to mitigate and remediate these limitations. A proactive approach to addressing these issues can help ensure the integrity and quality of EHR data and the appropriateness of their use in clinical studies. **(Gut Liver 2024;18:201-208)**

Key Words: Electronic health records; Cohort studies; Data accuracy; Bias

INTRODUCTION

Electronic health records (EHRs) have been adopted in over 90% of hospitals and office-based practices in the United States.[1-3] This digitalization of the health care system has led to increased research using the EHR as a data source. Unlike population-based registries and large administrative databases, the EHR does not require intense additional resources to develop or maintain beyond those during the clinical data imputation providing a source of large volumes of up-to-date, longitudinal, real-world data.[4] Data extraction from the EHR may be performed immediately, in contrast to delays often seen with claims databases. Clinical notes provide a level of detail that may not be available in conventional databases and registries.

Despite these multiple advantages, it is important to understand the EHR was not originally designed for research purposes, but rather to serve as a data repository and billing system. This leads to multiple limitations and challenges of the EHR relating to clinical research, such as data availability, biases and errors relating to available data and their use in clinical research.[5,6]

The successful and appropriate use of EHR data for research purposes entails consideration of the quality of the data, accounting for potential errors and biases, and developing effective strategies to mitigate these limitations. Development of strategies to address these limitations will allow results of EHR-based research to be validated and generalizable to the population of interest.[1,7,8]

In this review article, we present a comprehensive description of the limitations, challenges and opportunities relating to EHR-based research. We discuss the types of

data available in the EHR and challenges pertaining to data quality including errors and biases. We then examine and identify potential sources of such biases and errors and potential strategies including currently available approaches to mitigate these concerns.

DATA IN THE EHR

The EHR includes both structured and unstructured data.[9] Structured EHR data refer to standardized and organized data fields with limited and discrete outcomes (Table 1). Examples may include sociodemographic data or data obtained during medical encounters (e.g., medications, diagnosis codes). Data stored in a structured format allows for easy retrieval and analysis but do not provide insight into the overall clinical context. In contrast, unstructured data refer to the free-text documents and clinical narrative notes found in nursing and physician notes, discharge summaries, procedures, imaging, and pathology notes (Table 2). Unstructured data contain details relating to patients' symptoms, history and other elements not captured by coded organized data. While this level of detail is what researchers need for accurate data, given their unstandardized format, unstructured data can be challenging to extract and analyze. Technologies such as natural language processing (NLP) and machine learning models may be used to retrieve this type of data.[10] Both structured and

Table 1. Variables of Interest in Structured Data

Variable	Data source	Data propagation	Potential types of error or bias	Relative likelihood of error or bias	Change over time
Sex	Patient	Auto-propagate	Misclassification	Low	Static
Race/ethnicity*	Patient	Auto-propagate	Misclassification	Low	Static
Vital signs	Provider's assistant	NA	Measurement error Recording error Selection bias	Low	Moderate
Height, weight, BMI	Patient Provider's assistant	Auto-propagate	Reporting bias Selection bias Measurement error Time-dependent	Low	Moderate
Medical history*	Patient Provider	Auto-propagate	Selection bias Recall bias	Medium	Moderate
Family history*	Patient Provider	Auto-propagate	Selection bias Recall bias	Medium	Moderate
Problem list	Patient Provider	Forward-propagate	Systematic error Recall bias	High	Dynamic
Medication list	Patient Assistant Provider	Forward-propagate	Systematic error Recall bias	High	Dynamic
Smoking/alcohol history*	Patient	Auto-propagate	Reporting bias Recording error	High	Dynamic
Visit diagnoses	Provider	NA	Misclassification	Medium	Dynamic
Laboratory values	Automatic entry	NA	Selection bias	Low	Dynamic

BMI, body mass index; NA, not available.
*Variables may be recorded as structured or unstructured data.

Table 2. Variables of Interest in Unstructured Data

Variable	Data source	Data propagation	Potential types of error or bias
Race/ethnicity*	Patient	Auto-propagate	Reporting bias
Symptoms	Patient	NA	Recall bias
Family history*	Patient	Auto-propagate	Recall bias
Medical history*	Patient	Auto-propagate	Reporting error
Imaging	Provider (auto/template)	Auto-propagate	Reporting error
Procedures	Auto (auto/template)	Auto-propagate	Reporting error
Pathology	Auto (auto/template)	Auto-propagate	Reporting error

NA, not available.
*Variables may be recorded as structured or unstructured data.

unstructured data are imperfect with many limitations pertaining to quality and accuracy, ranging from selective data entry to variability in practice and documentation.[11] Hence, it is important to understand potential errors and biases pertaining to EHR data use.

TYPES OF BIAS IN EHR DATA

Before reviewing potential errors and biases in EHR-based research, we present a brief definition of these terms. Error is defined as the difference between the true value of a measurement and the recorded value of a measurement that can occur during data collection or analysis. Random errors occur because of sampling variability that could be related to environmental conditions, human performance or equipment restrictions. Random errors decrease as sample size increases.[12] Systematic error or bias refers to deviations that are not due to chance alone. They can be introduced at any point in the study and are not a dichotomous variable. In other terms, the degree of bias present matters more than its overall presence or absence.[12] We now discuss the potential biases in EHR-based research.

1. Information bias

Information bias occurs when data are inaccurate because of missing input/results, measurement, or recording errors. A measurement error is the difference between a measured value and its true value. It includes random and systematic errors. Random error is caused by any factor that randomly affects the measurement of the variable across the sample, while systematic error is caused by a factor that systematically affects all measurements of the variable across the sample. A recording or data entry error refers to a inaccuracies in recording a health measurement. Generally recording errors are believed to be random, hence not considered true bias.[13,14]

Information biases include recall, reporting and misclassification biases. Recall and reporting bias result from differential accuracy in recall or reporting of an exposure or outcome respectively. To assess for the presence of such biases, it is important to compare the reported outcomes and analyses with the original study protocol or registration.

Misclassification bias is a type of information bias and refers to an incorrect recording of either an exposure or outcome, and can occur in two forms: differential and non-differential.[15] Non-differential misclassification is when the data entry error is random and not related to a specific factor, and hence would not systematically over or underestimate results (e.g., blood pressure). On the other hand, a differential misclassification can lead to over or underestimating the accuracy or severity of illness. An example would be diagnostic ICD codes entered for the purpose of billing and higher reimbursement or behavioral history related to substance use disorders[16] while patients tend to underreport substance use introducing a systematic bias.[12,14]

2. Selection bias

Selection bias occurs when the study population in the EHR does not adequately represent the intended population of interest.[17,18] Access to care and entrance into a healthcare setting is complex and often influenced by medical insurance.[19,20] In addition, multiple factors such as geography, care setting and offered services available at one particular health system may influence patients included in the EHR which may affect the representativeness of the study population and, therefore, the generalizability of the study findings. To assess for selection bias, it is important to compare the characteristics of the study population with those of the general population or other relevant populations.

Informed presence bias, that may exacerbate selection bias, occurs when only patients with adequate access to care may have undergone testing to establish a diagnosis.[21] In particular, underserved populations may be poorly represented in the EHR, due to poor access, utilization and fragmented care.[1,14,16] Thus, differential patient participation is another contributor to informed presence bias. If an investigator undertaking an EHR-based study elects to only include patients with sufficient data, the approach may introduce a bias towards sicker patients.[22]

3. Ascertainment bias

Ascertainment bias results from data acquisition due to clinical need. Practice-based evaluation differences with regards to extent of social and behavioral history contributes to such biases.[23] Ascertainment bias also occurs when differential measurement methods are applied to the group of interest such as the use of dot phrases and templates, which may influence the data obtained from patients.

SOURCES OF BIAS AND ERROR

To effectively mitigate bias, one must understand potential sources of bias and error when using EHR data for clinical research.[24] Some factors that may contribute to bias include missing data, data entry errors, patient compliance, and changes in patient status over time that are not reflected in the EHR.

1. Missing data

Data in the EHR only includes encounters performed within the health system. These may include services, tests and test results, procedures, and treatments. Patients may seek health care at more than one system, depending on multiple factors related to individual preferences such as geography and existing relationships with health care providers. In addition, the specific medical issue, urgency, chronicity of symptoms, and time of onset may influence access to health care and availability of providers to assess and treat the issue. Any care outside the health system may not be included in the EHR. This results in censoring: left censoring refers to the outcome of interest occurring before the start of the study and right censoring denotes an unobserved event or loss to follow-up at the time of or after study completion. Censoring is especially significant for studies assessing outcomes following hospitalizations or survival analyses.[25]

2. Data entry

Multiple methods of data entry may contribute to error in the EHR. Frequently used EHR templates include automated data entry such as medications and problem lists, potentially "forward-propagating erroneous data."[24] Similarly, the provider practice of "copy and paste" may also perpetuate outdated or incorrect data. Providers with busy clinical practices may provide more limited documentation, compared to more highly resourced providers with nursing, scribes or other support staff. Finally, billing requirements may influence provider behavior and promote attention to certain fields necessary for billing.

3. Patient adherence and compliance

Lack of patient adherence and compliance may serve as a source of measurement error and bias. For example, prescriptions reflect the orders written by providers, but not necessarily patient compliance. Adherence to and compliance with healthcare recommendations is a multifactorial process related to patient and physician-related factors and may be further complicated by the type of encounter.[26] Previous studies have demonstrated that follow-through on provider recommendations is significantly better with in-person encounters, compared to telehealth.[27] In addition, concordance between providers and patients with respect to language and culture (cultural sensitivity) influence patient uptake of provider recommendations.[28]

4. Changes over time

Time is an essential yet complex element in the EHR; potential considerations range from the time a health system adopted EHR, to date of disease onset, to treatment duration. As hospitals and health systems merge, create new partnerships or acquire new facilities, the EHR composition changes which may influence data captured over time. Date of disease onset is frequently a necessary variable to identify a cohort of interest; however, accurate identification remains challenging, as date of diagnosis and time of entry of an individual into the EHR may not align. Medication exposure and treatment duration are other important variables which do not exist as structured data in the EHR but are represented by proxy measures such as physician orders for a prescription. In particular, medication and problem lists are highly time-dependent and may be especially prone to systematic error.

Table 3. Best Practices: Use of Electronic Health Record Data In Clinical Research

Challenge	Approach
Evaluate population of interest	Evaluate representativeness of study population with respect to target population
Assess feasibility and accuracy of measuring outcome, exposure, and confounder variables	Ensure that outcome measurement mirrors outcome of interest Choose times for dynamic variable
Evaluate quality of data	Assess data missingness; report missing values Evaluate reason for missing data Compare cohort with complete vs incomplete data Confirm data missingness is random If data missingness is not random, assess for systematic error or bias
Assess for presence of bias, error and confounding	Quantitative bias analysis Evaluation of results
Provide context for results	Compare results with those published in medical literature
Address missing data	Imputation Multiple imputation Inverse proportional weighting Natural language processing
Validate results	Sensitivity analysis Internal validation External validation

ASSESSMENT FOR PRESENCE AND DEGREE OF ERROR AND BIAS

It is important to understand that systematic error and bias are not reduced with the use of large data and that assessing for the presence and degree of error is critical to interpretation of EHR-based research (Table 3).

1. Assessment of data quality and representativeness

When assessing for data quality, two main factors need to be considered: data representativeness and availability.[1,29] When contemplating using EHR data, one must ensure the population of interest is available and representative of the target population.[30] This could be conducted by a preliminary assessment of sociodemographic data. An evaluation of the approximate duration and density of relevant data in the EHR may also be needed. Comparing an EHR data sample to an external data source could be considered. If selection bias is suspected, one can then employ inverse probability weighting.[1,31]

Another important factor is data availability. The EHR was not originally designed for research purposes but to optimize billing, maintain clinical records and scheduling.[1,8] Recently, techniques such as NLP have been employed to capture details from clinical free-text notes. Missing data can lead to information bias and confounding. It is, therefore, important to assess missing data in both outcome and predictor variables and determine if they are missing at random or systematically.[21,32]

2. Statistical analyses

Several statistical methods help estimate bias magnitude and direction. Quantitative bias analysis may be performed in the design phase of the study to assess whether missing data is random or indicative of inclusion, misclassification or selection bias.[33] This will help investigators understand the data and research environment and mitigate potential biases before the analysis phase.[34] Quantitative bias analysis entails identifying potential sources of bias, estimating their magnitude and direction using previous literature or statistical methods, and incorporating those parameters into the analysis. Inter- or intra-observer variability for repeat measurements can be assessed using kappa coefficients. Bias should be evaluated by race, ethnicity, gender, and across time to ensure lack of unrecognized bias in different groups.[35]

One may also evaluate multiple approaches and select the best analysis method.[36] A selective approach may potentially produce higher quality data, but can be associated with the highest selection bias. In contrast, a common data approach (most inclusive) may produce lower quality data, but be associated with information/misclassification bias. A "best available data" approach may allow for a compromise between the competing demands of selection and inclusivity.

Assessing for confounding from missing measurements can be considered. In one study, an NLP-driven approach to identify potential confounders was described.[37] NLP was used to process unstructured data from clinical notes and creates covariates to supplement traditional known covariates. The dataset was augmented with NLP-identified covariates. This approach reduced selection bias and results aligned with those obtained from randomized controlled trials (RCTs).[37]

3. Evaluation of results in the context of medical literature

Preliminary results such as sociodemographic characteristics or median survival should be compared with expected outcomes as found in the literature. For instance, the incidence or prevalence of disease in an EHR can be compared to known population values such as Surveillance, Epidemiology, and End Results data. Results from comparative effectiveness studies should be compared to those available from randomized controlled studies.[18]

APPROACHES TO MITIGATE ERROR AND BIAS

Multiple approaches have been described to mitigate error and bias in EHR-based research. We present the most commonly described strategies currently in use and potential consequences of such approaches.

1. Addressing missing data

Individuals with missing data are usually addressed by excluding them from the study, which can lead to a potential loss of study power if a large portion of the population of interest is excluded, and to biased results.[32,38] The risk of bias largely depends on whether data is missed completely at random or systematically (at random or not at random).[38] If the data is missing completely at random, imputation and inverse proportional weighting can be used to adjust for the selection bias. Imputation is frequently performed and may include imputation from observed values (mean) or the last measured value (last value carried forward). However, this method does not account for the uncertainty around the missing value and may introduce systemic bias. If missingness is not at random, multiple imputation may better account for the uncertainty around missing data; this technique creates multiple imputed da-

tasets and combines results obtained from each of those sets.[29,32]

Another method to address missing data is to supplement EHR data with external data such as registries, intervention trials, or community health clinics.[21] Obtaining dispensing claims in pharmacy level data for medications can also be used.[24,39] Access to high-quality external data or summary statistics have enabled investigators to develop statistical methods that account for simultaneous misclassification and selection biases.[40]

The use of NLP to retrieve data from unstructured data is being increasingly used.[15] NLP offers the benefit of assessing unstructured data and organizing them into more discrete variables and concepts, but may also introduce systematic errors. In a study where NLP was applied to recover vital signs from free-text notes missingness of vital signs were reduced by 31% and the recovered vital signs were highly correlated with values from structured fields (Pearson r, 0.95 to 0.99).[41]

2. Validating results/models

1) Validation study

For studies involving the development of clinical prediction models using artificial intelligence, machine learning or regression-based models, results must first be internally validated by stratifying the cohort into a development and validation set. The model quality and performance can be evaluated by metrics such as area under the receiving operating characteristics curve, area under the precision-recall curve, sensitivity, positive predictive value, negative predictive value, c-statistic, and r-coefficient. This is followed by external validation of the prediction model performance, a critical step to ensure that the results are generalizable to populations not involved in the model development process.[42]

2) Sensitivity analysis

Sensitivity analyses should be performed to confirm robustness of results and ensure that the results (e.g., model performance) hold across a range of values. This approach can evaluate how different values of independent variables affect a particular dependent variable under a given set of assumptions.[43] In particular, sensitivity analysis can assess whether alteration of any of the assumptions will lead to different results in the primary endpoints. If the results in the sensitivity analysis are consistent with the results in the primary analysis, then it increases confidence that assumptions that are inherent in modeling and the EHR data (e.g., missing data, outliers, baseline imbalance, distribution assumptions, and unmeasured confounding) had negligible impact on the results. It is advisable for sensitivity analyses to be considered and reported in EHR-based studies.[44]

3. Addressing confounding

Multiple methods have been described to address confounding in EHR-based studies.[45-48] Using a traditional adjustment for measured confounders by using propensity scores in the main cohort unmeasured confounding by estimating additional propensity scores can be addressed in a validation study.[45] Regression calibration can be applied to adjust regression coefficients, leading to a calibration of propensity scores. A Bayesian nonparametric approach for causal inference on quantiles has also been described to adjust for bias in the setting of many confounders.[48]

A recently described use of NLP is to address and uncover potential confounders.[37] An NLP-based framework to uncover potential confounders in unstructured data from free-text notes was developed and hazard ratios with and without confounding covariates was compared with previously published RCTs.[37] With the additional confounding covariates, the estimated hazard ratios were able to be shifted toward the direction of the results obtained in RCTs. Inverse proportional weighting is another approach to address confounding: after identifying confounding variables, inverse proportional weights are assigned to each observation and incorporated in the statistical analysis. This allows adjusting for multiple exposure confounders.[49]

CONCLUSIONS

With the growing interest in using EHR data for observational cohort studies, it is important to recognize that large volume and longitudinal data do not necessarily increase data validity and study power but can incorporate significant biases and potentially decrease the validity of a study. Missing data is the most important source of error, while selection, information, and ascertainment biases may substantially influence available data and measured outcomes. These errors and biases may exist at the planning, data extraction, analysis, or result interpretation phases of a study. Multiple techniques assist in identifying the magnitude and direction of bias. Statistical techniques and NLP-based approaches may assist in mitigating biases and confounders. The EHR could be a valuable, high-quality source of data for observational and experimental studies; however, researchers must remain aware of the inherent limitations of EHR data, and apply the different approaches described to mitigate those challenges.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGEMENTS

This study was supported in part by NIH K08 AA028794 (N.W.); R01 GM119174; R01 DK113196; P50 AA024333; R01 AA021890; 3U01AA026976-03S1; U01 AA 026976; R56HL141744; U01 DK061732; 5U01DK062470-17S2; R21 AR 071046; R01 CA148828; R01CA245546; R01 DK095201 (S.D.).

ORCID

Michelle Kang Kim https://orcid.org/0000-0001-5285-8218
Carol Rouphael https://orcid.org/0000-0003-3492-2073
John McMichael https://orcid.org/0000-0002-4592-0500
Nicole Welch https://orcid.org/0000-0002-4655-5146
Srinivasan Dasarathy https://orcid.org/0000-0003-1774-0104

REFERENCES

1. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. BMC Med Res Methodol 2021;21:234.
2. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc 2017;24:1142-1148.
3. HealthIT.gov. Office-based physician electronic health record adoption [Internet]. Washington, DC: HealthIT.gov [cited 2019 Jan 15]. Available from: https://www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption
4. 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;37:61-81.
5. Haneuse SJ, Shortreed SM. On the use of electronic health records. In: Gatonsis C, Morton SC, eds. Methods in comparative effectiveness research. Boca Raton: CRC Press, 2017:449-482.
6. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. JAMA Netw Open 2021;4:e210184.
7. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017;106:1-9.
8. Hersh WR. The electronic medical record: promises and problems. J Am Soc Inf Sci 1995;46:772-776.
9. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 2014;21:801-807.
10. Coleman N, Halas G, Peeler W, Casaclang N, Williamson T, Katz A. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. BMC Fam Pract 2015;16:11.
11. 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 2018;20:e185.
12. Pannucci CJ, Wilkins EG. Identifying and avoiding bias in research. Plast Reconstr Surg 2010;126:619-625.
13. Bower JK, Patel S, Rudy JE, Felix AS. Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise. Curr Epidemiol Rep 2017;4:346-352.
14. Bots SH, Groenwold RH, Dekkers OM. Using electronic health record data for clinical research: a quick guide. Eur J Endocrinol 2022;186:E1-E6.
15. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016;23:1007-1015.
16. Jha AK, DesRoches CM, Campbell EG, et al. Use of electronic health records in U.S. hospitals. N Engl J Med 2009;360:1628-1638.
17. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15:615-625.
18. Arterburn D, Aminian A, Nissen S, Schauer P, Haneuse S. Bias in electronic health record-based studies: seeing the forest for the trees. Diabetes Obes Metab 2021;23:1692-1693.
19. Centers for Disease Control and Prevention (CDC). Vital signs: health insurance coverage and health care utilization: United States, 2006--2009 and January-March 2010. MMWR Morb Mortal Wkly Rep 2010;59:1448-1454.
20. Miller S, Wherry LR. Health and access to care during the first 2 years of the ACA Medicaid expansions. N Engl J Med 2017;376:947-956.
21. Phelan M, Bhavsar NA, Goldstein BA. Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference. EGEMS (Wash DC) 2017;5:22.
22. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC

Med Inform Decis Mak 2014;14:51.
23. Hollister B, Bonham VL. Should electronic health record-derived social and behavioral data be used in precision medicine research? AMA J Ethics 2018;20:E873-E880.
24. Young JC, Conover MM, Funk MJ. Measurement error and misclassification in electronic medical records: methods to mitigate bias. Curr Epidemiol Rep 2018;5:343-356.
25. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013;51(8 Suppl 3):S30-S37.
26. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018;25:1080-1088.
27. Shah VV, Villaflores CW, Chuong LH, et al. Association between in-person vs telehealth follow-up and rates of repeated hospital visits among patients seen in the emergency department. JAMA Netw Open 2022;5:e2237783.
28. Cano-Ibáñez N, Zolfaghari Y, Amezcua-Prieto C, Khan KS. Physician-patient language discordance and poor health outcomes: a systematic scoping review. Front Public Health 2021;9:629041.
29. Li J, Yan XS, Chaudhary D, et al. Imputation of missing values for electronic health record laboratory data. NPJ Digit Med 2021;4:147.
30. Bower JK, Bollinger CE, Foraker RE, Hood DB, Shoben AB, Lai AM. Active use of electronic health records (EHRs) and personal health records (PHRs) for epidemiologic research: sample representativeness and nonresponse bias in a study of women during pregnancy. EGEMS (Wash DC) 2017;5:1263.
31. Goldstein ND, Kahal D, Testa K, Burstyn I. Inverse probability weighting for selection bias in a Delaware community health center electronic medical record study of community deprivation and hepatitis C prevalence. Ann Epidemiol 2021;60:1-7.
32. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.
33. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol 2014;43:1969-1985.
34. Fox MP, Lash TL. Quantitative bias analysis for study and grant planning. Ann Epidemiol 2020;43:32-36.
35. Estiri H, Strasser ZH, Rashidian S, et al. An objective framework for evaluating unrecognized bias in medical AI models predicting COVID-19 outcomes. J Am Med Inform Assoc 2022;29:1334-1341.
36. Hubbard RA, Lett E, Ho GY, Chubak J. Characterizing bias due to differential exposure ascertainment in electronic health record data. Health Serv Outcomes Res Methodol 2021;21:309-323.
37. Zeng J, Gensheimer MF, Rubin DL, Athey S, Shachter RD. Uncovering interpretable potential confounders in electronic medical records. Nat Commun 2022;13:1014.
38. Little R, Rubin D. Statistical analysis with missing data. 2nd ed. New York: Wiley, 2002.
39. Fischer MA, Stedman MR, Lii J, et al. Primary medication non-adherence: analysis of 195,930 electronic prescriptions. J Gen Intern Med 2010;25:284-290.
40. Beesley LJ, Mukherjee B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. Biometrics 2022;78:214-226.
41. Khurshid S, Reeder C, Harrington LX, et al. Cohort design and natural language processing to reduce bias in electronic health records research. NPJ Digit Med 2022;5:47.
42. Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. Diagn Progn Res 2022;6:24.
43. Thabane L, Mbuagbaw L, Zhang S, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med Res Methodol 2013;13:92.
44. Parpia S, Morris TP, Phillips MR, et al. Sensitivity analysis in clinical trials: three criteria for a valid sensitivity analysis. Eye (Lond) 2022;36:2073-2074.
45. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol 2005;162:279-289.
46. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 2009;20:512-522.
47. Zigler CM, Dominici F. Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects. J Am Stat Assoc 2014;109:95-107.
48. Xu D, Daniels MJ, Winterstein AG. A Bayesian nonparametric approach to causal inference on quantiles. Biometrics 2018;74:986-996.
49. Buchanan AL, Hudgens MG, Cole SR, Lau B, Adimora AA; Women's Interagency HIV Study. Worth the weight: using inverse probability weighted Cox models in AIDS research. AIDS Res Hum Retroviruses 2014;30:1170-1177.