Leveraging electronic health records for data science: common pitfalls and how to avoid them

Christopher M Sauer, Li-Ching Chen, Stephanie L Hyland, Armand Girbes, Paul Elbers, Leo A Celi

Analysis of electronic health records (EHRs) is an increasingly common approach for studying real-world patient data. Use of routinely collected data offers several advantages compared with other study designs, including reduced administrative costs, the ability to update analysis as practice patterns evolve, and larger sample sizes. Methodologically, EHR analysis is subject to distinct challenges because data are not collected for research purposes. In this Viewpoint, we elaborate on the importance of in-depth knowledge of clinical workflows and describe six potential pitfalls to be avoided when working with EHR data, drawing on examples from the literature and our experience. We propose solutions for prevention or mitigation of factors associated with each of these six pitfalls—sample selection bias, imprecise variable definitions, limitations to deployment, variable measurement frequency, subjective treatment allocation, and model overfitting. Ultimately, we hope that this Viewpoint will guide researchers to further improve the methodological robustness of EHR analysis.

Introduction and context: use of data from electronic health records

Electronic health records (EHRs) are increasingly used to survey population health, develop classification and prediction models for decision support, identify optimal treatment policies, or even simulate randomised clinical trials. Compared with alternative data sources and study approaches, such as cohort studies or randomised controlled trials, the main advantage of EHR analysis is the use of data that are already captured, thus reducing administrative efforts, costs, and sample selection bias. EHRs are more representative of the total target population and less subject to inclusion bias than randomised controlled trials, because EHR data are obtained from all individuals who interact with health systems. Although important epidemiological considerations remain, such as those discussed in this Viewpoint, EHRs include more comprehensive data from patients as they evolve over time and provide access to large dataset sizes, especially when EHR analysis involves large integrated health-care systems or a network of health-care providers who use interoperable information systems. Ultimately, large sample sizes can provide increased statistical power to do subgroup analyses and reduce the risk of type II errors.

Importance of methodological robustness

Although EHR data analysis offers promising research opportunities, the applicability and robustness of EHR research are often limited by methodological shortcomings inherent to the complexity of EHR data.

First, compared with clinical repositories and clinical trials, the likelihood of variation in data collection for EHRs is increased because many variables are measured by use of different technologies and sampling frequencies. Second, the retrospective nature of most EHR data analyses means that cohorts, exposures, and outcomes are defined retrospectively. This retrospective approach allows for such definitions to be adjusted in light of analysis outcomes, resulting in potentially misleading findings, because of multiple hypothesis testing and without appropriate statistical adjustment. Finally, the type and quality of data, including which patients are included in the analysis, are heavily dependent on clinical practice patterns, which introduce sample selection biases that lead to spurious associations. One should keep in mind that EHR data are not primarily collected for research purposes; instead, they store all patients’ information collected during clinical care, and, in some cases, serve an administrative function such as billing. Therefore, analysis of EHR data should not be undertaken without the involvement of investigators with specialist knowledge of the domain of interest, spanning from care practices to data processing.

Goals of this Viewpoint

Over the past two decades, with ever increasing use of EHR data, previous investigations and our research have observed that issues arising during clinical care (eg, from minority ethnic groups seeking health care less frequently) are also reflected in the EHR data, thus introducing bias in EHR studies. We believe that an integrated and comprehensive discussion of how clinical workflows and information system design affect EHR data analyses is still missing in the literature. In this Viewpoint, we extend these concepts by integrating machine learning, epidemiological, and statistical considerations with important clinical considerations, by reviewing the literature and case examples. Although most pitfalls we present are mainly relevant to machine learning models, others are also crucial for all types of EHR studies.

In this Viewpoint, we identify six common clinical and methodological pitfalls using a mixed approach, both relying on expert opinion and a literature review. We cross-checked our literature review and references of the identified publications to include additional relevant publications. Inclusion of papers was at the authors’ discretion (CMS, SLH, and LAC). By discussing these six pitfalls of crucial importance, we hope to guide
researchers to avoid repeating the same errors in future analyses, and we hope that the solutions we provide will improve the scientific robustness of descriptive and predictive models that use EHR data. Although some points we present here are elaborations or extensions of key concepts that have been previously published in reporting guidelines such as the Strengthening the Reporting of Observational studies in Epidemiology checklist,14 we also describe additional pitfalls and concepts. Furthermore, the added value of this Viewpoint lies in the provision of actionable solutions.

Pitfalls and solutions
Arguably most mistakes are not made on purpose or through negligence, but by not knowing what one does not know. Especially when trying to build advanced statistical models, in-depth knowledge of how data are collected and how decisions are made is crucial, such as when patients are admitted to, for example, hospital or an intensive care unit (ICU), how treatment decisions are made, or when patients should be discharged from an ICU or from hospital. Although non-medical researchers might obtain some insights into these processes from the internet or from the media, we recommend involving clinicians, nurses, or other relevant health-care providers throughout the lifecycle of the research project.15 In several previous analyses, the influence of hospital-specific practices was not incorporated, ultimately leading to issues with the external (and sometimes also internal) validity of the studies16–18 and highlighting the importance of domain expertise on local health-care unit policies in EHR research.19 Protocols that vary among hospitals can establish if, when, and how some data are collected. For example, an atypical blood test result can only be retrieved from the EHR if the patient had a blood test. Thus, local protocols and diverse testing frequencies for different patient populations can be associated with the identification of patients with atypical results. This issue of missing data is a very common, yet overlooked, problem in research and leads to sample selection bias that is seldom adjusted for.

When analysing EHR data collected from patients in ICUs, the analysis is affected by the criteria for admission to the ICU, which vary among different ICUs and even within the same hospital depending on circumstances.20–24 Such changing circumstances became obvious during the COVID-19 pandemic, when demand for ICU beds exceeded capacity.25 Similarly, many other clinical decisions are susceptible to subjectivity of health-care providers, such as when a patient is ready for discharge (selective censoring),26 or when a patient should be readmitted from the ward to the ICU.27 In most of these situations, factors related to health-care providers, such as clinical experience and cultural differences, or current events in general, such as the pandemic, can affect triage and other decisions. To address the aforementioned issues, we suggest always discussing design choices and study assumptions with clinicians or other health-care providers who are knowledgeable of local protocols. Furthermore, causal inference frameworks28 should be incorporated if studies aim to analyse or predict outcomes that result from treatment decisions (eg, which patients should be offered renal replacement therapy). We have listed the identified pitfalls and outlined potential solutions (figure). The pitfalls are structured according to the analysis stage in which they most likely occur.

Pitfall 1: sample selection bias
Cohort building is a challenging step in data analysis because it requires translation from diverse case definitions into specific data criteria. The quality of the definition used in the literature to correctly identify cases, and the interpretation and translation of these definitions by researchers affect the composition of the cohort. Crucially, sample selection bias can occur if these case definitions are overly restrictive, thus potentially excluding a subgroup, or too general, thus increasing the number of falsely identified patients in the cohort.29–30

One prominent example of definitions being too unspecific is seen in studies using different methods to identify patients with sepsis, including the quick Sequential Organ Failure Assessment (qSOFA) score,31 codes proposed by Martin and colleagues32 (known as Martin methodology), codes proposed by Angus and van der Poll33 (Angus methodology), or the systemic inflammatory response syndrome score.34 Although use of these methods as a surrogate for sepsis is established,35–37 such methods lack specificity. Meanwhile, performance to identify patients with sepsis in various situations has repeatedly been shown to be variable and moderately discriminatory, reflected in areas under the receiver operating characteristic curve (AUROC) values between 0·6 and 0·8 for qSOFA scores of 2 or higher.38–40

Furthermore, misclassification intrinsic to disease definition can occur, such as when applying the gold-standard definition known as Sepsis-3.41 One crucial part of the Sepsis-3 definition is suspicion of infection, for which more inclusive definitions of what constitutes a suspected infection in place of restrictive ones (eg, whether a positive nitrite urine test is sufficient or whether also other signs of tissue invasion have to be present) could substantially affect cohort size and model performance.42 The poor performance of the Epic Sepsis Model is an example of why cohort definitions and external validation are important factors to consider in a study.43 Notably, even when algorithms are used to define syndromes such as sepsis, retrospective review of patients considered to have sepsis might show serious misclassification.44–46

Another example of definitions being too general is associated with mortality prediction: constructed cohorts used in different studies have been observed to be
variable, despite all publications using the same dataset (Medical Information Mart for Intensive Care III) and investigating the performance of the same model (ie, mortality prediction). Sources of variation in the inclusion and exclusion criteria included age restrictions (eg, excluding individuals younger than 18 years), the exclusion of individuals with multiple ICU stays, or the requirement of particular measurements, such as those of infection markers.

On the basis of the aforementioned examples, we suggest always using the definitions that have been extensively validated but appreciating that even these definitions are prone to error and not fixed. Furthermore, if reported, model performance should be compared if reported, model performance should be compared to existing algorithms, whether expert opinion-based or machine learning-based; if algorithms based on machine learning are used for comparison, differences in definitions should also be explored. We suggest explicitly describing the definitions used for cohort identification, explaining potential limitations, and discussing potential effects on the robustness of the analyses. If no gold-standard definitions are available, we suggest performing sensitivity analyses and checking which approach succeeds best in identifying the target population.

Pitfall 2: imprecise variable definitions
After cohort building, imprecise variable definitions is the next important pitfall. As is the case for cohort building, definitions should be meticulously inspected because their sensitivity and specificity have substantial influence on whether the analyses truly reflect real-world practice.

When deciding on the outcome measure, one should appraise the epidemiological implications and the clinical context. A common example is the question of whether to use in-hospital events (eg, hospital mortality) or a fixed timepoint (eg, 28-day mortality). Generally, in-hospital events should be used when studying effects associated with direct hospital interventions, such as the effect of prophylactic heparin use on incidence of venous thromboembolism in patients treated in hospital. By contrast, when interested in the occurrence of postsurgical thrombosis, a fixed timepoint should be used because length of stay in hospital is associated with risk of thrombosis.

Generally, we recommend obtaining advice on definitions from a multidisciplinary team including patients and people with relevant subject matter knowledge, such as clinicians, epidemiologists, statisticians, and social scientists.

Pitfall 3: limitations to deployment
A major challenge to real-world use and deployment of machine learning or other predictive models developed from EHR studies is that data available in EHRs cannot be easily translated into clinical practice. This issue arises if data structure and availability of time-stamped results in EHRs differ from real-world practice. For instance, during retrospective analysis, time-stamped data might be assumed to be available at the time of measurement and analysed as such. However, these data might not be realistically available to clinicians or other decision makers at the expected time. Such discrepancy can, for example, occur when results of blood tests or blood cultures are used: they might be included in the analysis with the timestamp of the registration of the blood draw, and not the time they became available to the clinicians.

Another example is exclusion of patients with very long hospital stays—information that is not available early in
the patient’s hospital stay (ie, when most algorithms are deployed). Furthermore, International Classification of Diseases codes are typically assigned to patients after hospital discharge or death and with varying timestamps for their occurrence; therefore, their use would not be appropriate for models using hospital admission as a baseline because of the temporal difference. In many of the aforementioned situations, model performance might be overestimated and might not be replicable in real-world applications.

A separate but related issue is the leakage of data, which refers to the appearance of training data in the test datasets. This leakage can often occur if multiple admissions from the same patients are used or if time-series data from one patient are not isolated to the training or test set but appear in both. Leakage of data is an important methodological flaw that can result in overestimation of model performance, thus limiting performance and usefulness of the model in clinical practice.

Challenges associated with deployment can be avoided by understanding when data become available in real time and by selecting features on the basis of their data availability to physicians. Furthermore, time-series data and data from multiple admissions to, for example, hospital or ICU should be randomly assigned only to either the training or test dataset to avoid leakage of patient data from the training to the test dataset.

**Pitfall 4: variable measurement frequency**

Another often overlooked issue is the intrinsic association between frequency of measurements and severity of disease. Clinicians typically order additional laboratory measurements or review recordings of vital signs more frequently if patients are unstable. This association between frequency of measurements and disease severity is important when building predictive or descriptive models. For example, model validity is affected when patients with a substantial proportion of missing (ie, not performed) values are excluded from the study. This exclusion often results in a biased model that might potentially overestimate severity of illness, thus constituting a form of sample selection bias (pitfall 1). Importantly, imputation with mean or median values from the cohort is not a solution to compensate for missing values because measurement frequency is not a random occurrence. This example shows how routinely collected data implicitly encode clinicians’ judgement and might be highly variable across health-care providers. A model trained on such data can be susceptible to poor performance if variation in clinical practice is substantial and if clinician’s behaviour were to change.

Consequently, we recommend involving clinicians to identify circumstances in which variable measurement frequency could result in a biased analysis and statisticians to discuss appropriate epidemiological (eg, weighting) or statistical strategies (eg, multiple imputation, or removal of highly unreliable variables entirely).**Pitfall 5: subjective treatment allocation**

Causal inference studies typically aim to predict the unobserved, counterfactual treatment outcome enabling estimation of treatment effects. To quantify such effects, all factors associated with treatment allocation should be identified and corrected for. Although this approach has merit and has been shown to simulate randomised controlled clinical trials, it is not without limitations. The assumption for causal inference studies is that all factors influencing treatment allocation are observed. However, previous studies have shown that treatment allocation is subject to both differences in interphysician decision making (ie, different physicians prescribing different treatments in the same clinical setting) and intraphysician decision making (ie, biased on patient’s socioeconomic factors, such as race and ethnicity). Where possible, researchers should adopt causal frameworks when designing EHR-based studies to avoid introducing bias due to confounding. Before exploratory analyses, we recommend creating causal diagrams describing the team’s knowledge of the data generating process (eg, using directed acyclic graphs). These diagrams can identify missing confounders or other important factors that should be accounted for during the analysis. Adjusting for confounders is the basis of emulating randomisation, yet EHRs often do not capture relevant social determinants of health or clinical parameters. In this case, we recommend carefully considering whether a causal inference study is feasible at all. When in doubt, sensitivity analyses should be done to estimate the potential effect of unmeasured confounders on the basis of previous studies.

**Pitfall 6: model overfitting and reduced generalisability**

Differences among institutions and regions can be substantial and results might not be generalisable beyond the original data source. Notably, because of model overfitting, not all analyses and models need to be generalisable. Researchers should thus ask themselves if local practice patterns influence treatments or outcomes of interest. For example, if researchers within a hospital would like to establish how patient handover between services affects clinical outcomes, the best performing model could be developed without giving too much importance to issues arising from poor generalisability. However, if the intent is to address this topic beyond one hospital, external validation on another cohort should be done. Furthermore, differences between clinical settings in the prevalence of important variables, or their absence, should be highlighted and adjusted for where possible.

External validity is not a yes or no question but pertains to identifying which specific clinical contexts the analysis is relevant to. A causal diagram can be helpful to infer the generalisability of models, by making explicit which
associations in the data are likely to differ among institutions or regions.

Proper performance measures not affected by class imbalance should also be used to assess model performance. In case of imbalance in the frequency of a feature or outcome of interest, such as a very low hospital mortality for some medical conditions, the inclusion of metrics that are prone to class imbalance (e.g., accuracy) should be avoided as model performance measures. Furthermore, reporting only aggregate measures of discrimination (e.g., AUROC) might conceal reduced clinical usability because of, for example, not high enough sensitivity or specificity. Use of metrics based on a single operating point, such as the F1 score, which balances precision and recall, might be more informative. Ultimately, model assessment should be tailored to the intended use case of the system (e.g., screening vs treatment recommendation).

Conclusions

We identified six common methodological and clinical issues that limit robustness, validity, and reproducibility of studies leveraging EHR data. Key challenges are sample selection bias, imprecise variable definitions, limitations to deployment, absence of adjustment for the association between frequency of measurements and severity of disease, subjective treatment allocation, and reduced generalisability of findings. Although this list is not exhaustive and the suggested solutions to these issues are not universally applicable, we hope this Viewpoint will increase awareness of some crucial pitfalls associated with EHR data and encourage researchers to consider them when designing EHR-based studies.

Contributors

CMS, SLH, and LAC conceptualised the manuscript, outlined the study design, and were responsible for project administration. CMS, L-CC, and SLH wrote the original draft of the manuscript and were responsible for the literature search and data curation. CMS, L-CC, SLH, PE, and LAC did the analysis and investigation of the literature search results and curated the list of pitfalls. CMS designed the figure. AG, PE, and LAC contributed to the review and editing of the manuscript.

Declaration of interests

SLH is an employee of Microsoft Research (UK) and a board member of the non-profit organisation Association for Health Learning and Inference. All other authors declare no competing interests.

Acknowledgments

LAC is funded by the National Institutes of Health through the National Institute of Biomedical Imaging and Bioengineering (R01 grant EB017205).

References

1 Jacoba CMP, Celi LA, Silva PS. Biomarkers for progression in diabetic retinopathy: expanding personalized medicine through integration of AI with electronic health records. *Semin Ophthalmol* 2021; 36: 250–57.

2 Robles Arévalo A, Maley JH, Baker L, et al. Data-driven curation process for describing the blood glucose management in the intensive care unit. *Sci Data* 2021; 8: 80.

3 Sauer CM, Gómez J, Botella MR, et al. Understanding critically ill sepsis patients with normal serum lactate levels: results from US and European ICU cohorts. *Sci Rep* 2021; 11: 20076.

4 Komorowski M, Celi LA, Babawi O, Gordon AC, Faissal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* 2018; 24: 1716–20.

5 Tannen RL, Weiner MG, Marcus SM. Simulation of the Syt-Eur randomized control trial using a primary care electronic medical record was feasible. *J Clin Epidemiol* 2006; 59: 254–64.

6 Chari A, Romanus D, Palumbo A, et al. Randomized clinical trial representativeness and outcomes in real-world patients: comparison of 6 hallmark randomized clinical trials of relapsed/refractory multiple myeloma. *Clin Lymphoma Myeloma Leuk* 2020; 20: 8–17.

7 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: 366–52.

8 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013; 20: 144–51.

9 Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017; 106: 1–9.

10 Savitz TA, Savitz LA, Fleming NS, Shahli ND, Go AS. How much can we trust electronic health record data? *Healthc (Amst)* 2020; 8: 100444.

11 Fedor SL. Data quality in electronic health records research: quality domains and assessment methods. *Wost J Nurs Res* 2018; 40: 753–66.

12 Andra Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021; 375: n2281.

13 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.

14 von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PloS Med* 2007; 4: 247–51.

15 Gabbard J, Pajewski NM, Callahan KE, et al. Effectiveness of a nurse-led multidisciplinary intervention vs usual care on advance care planning for vulnerable older adults in an accountable care organization: a randomized clinical trial. *JAMA Intern Med* 2021; 181: 361–69.

16 He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019; 25: 30–36.

17 Harutyunyan H, Khachatryan H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Sci Data* 2019; 6: 96.

18 Beaulieu-Jones BK, Yuan W, Brat GA, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ Digit Med* 2021; 4: 62.

19 Bonomi S. The electronic health record: a comparison of some European countries. In: Ricciardi F, Harfouche A, eds. Information and communication technologies in organizations and society. Lecture notes in information systems and organisation, vol 15. Cham: Springer; 2016: 33–50.

20 Tansone V, Boudreau D, Ciccozzi M, et al. Ethical criteria for the admission and management of patients in the ICU under conditions of limited medical resources: a shared international proposal in view of the COVID-19 pandemic. *Front Public Health* 2020; 8: 284.

21 American Thoracic Society. Fair allocation of intensive care unit resources. *Am J Respir Crit Care Med* 1997; 156: 1282–301.

22 Curtis JR, Vincent JL. Ethics and end-of-life care for adults in the intensive care unit. *Lancet* 2010; 376: 1347–53.

23 Piers RD, Azoulay E, Ricou B, et al. Perceptions of appropriateness of care among European and Israeli intensive care unit nurses and physicians. *JAMA* 2011; 306: 2694–703.

24 Lapsley I, Melia K. Clinical actions and financial constraints: the limits to rationing intensive care. *Soc Med Health Illn* 2001; 23: 729–46.

25 Trentini F, Marziano V, Guzzetta G, et al. The pressure on health-care system and intensive care utilization during the COVID-19 outbreak in the Lombardy region of Italy: a retrospective observational study in 43 538 hospitalized patients. *Am J Epidemiol* 2021; 191: 37–46.

26 Thoral PJ, Formosa M, de Bruin DP, et al. Explainable machine learning on AmsterdamUMCdb for ICU discharge decision support: uniting intensivists and data scientists. *Crit Care Explor* 2021; 3: e0529.
27 Cosgrove CV, Celi LA, Sauer CM. Boosting clinical decision-making: machine learning for intensive care unit discharge. Ann Am Thorac Soc 2018; 15: 804–6.
28 Kleinberg S, Hripkas G. A review of causal inference for biomedical informatics. J Biomed Inform 2011; 44: 1002–12.
29 Natsanson C, Esposito CJ, Banks SM. The sirens’ songs of confirmatory sepsis trials: selection bias and sampling error. Crit Care Med 1998; 26: 1927–31.
30 Dueráñ HS, Seah C, Johnson JS, Huckins LM. Implicit bias of encoded variables: frameworks for addressing structured bias in EHR-GWAS data. Hum Mol Genet 2020; 29: R33–41.
31 Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016; 315: 762–74.
32 Martin GS, Mannino DM, Eaton S, Moss M. The epidemiology of sepsis in the United States from 1979 through 2000. N Engl J Med 2003; 348: 1546–54.
33 Angus DC, van der Poll T. Severe sepsis and septic shock. N Engl J Med 2013; 369: 840–51.
34 Bone RC, Balk RA, Cerra FB, et al. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Chest 1992; 101: 1644–55.
35 Zhao L, Zhao L, Wang YY, et al. Platelets as a prognostic marker for sepsis: a cohort study from the MIMIC-III database. Medicine (Baltimore) 2020; 99: e21511.
36 Ye W, Chen X, Huang Y, et al. The association between neutrophil-to-lymphocyte ratio and mortality in septic patients: a retrospective analysis of the MIMIC-III database. J Thorac Dis 2020; 12: 1843–55.
37 Zhou Y, Yang D, Fu Q, Chen T, Chen Y, Zhang C. Outcomes for patients with sepsis following admission to the intensive care unit based on health insurance status: a study from the Medical Information Mart for Intensive Care-III (MIMIC-III) database. Med Sci Monit 2020; 26: e924954.
38 Liu J, Jiang W, Wang L, et al. Usefulness of qSOFA and SIRS scores for detection of inceptip sepsis in general ward patients: a prospective cohort study. J Crit Care 2019; 51: 13–18.
39 Perman SM, Mikkelsen ME, Goyal M, et al. The sensitivity of qSOFA calculated at triage and during emergency department treatment to rapidly identify sepsis patients. Sci Rep 2020; 10: 20395.
40 Usman OA, Usman AA, Ward MA. Comparison of SIRS, qSOFA, and NEWS for the early identification of sepsis in the emergency department. Am J Emerg Med 2019; 37: 1490–97.
41 Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016; 315: 801–10.
42 Johnson AEW, Ahoab J, Raffa JD, et al. A comparative analysis of sepsis identification methods in an electronic database. Crit Care 2018; 46: 494–99.
43 Wong A, Otes E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021; 181: 1065–70.
44 Helfner AC, Horton JM, Marchick MR, Jones AE. Etiology of illness in patients with severe sepsis admitted to the hospital from the emergency department. Clin Infect Dis 2010; 50: 814–20.
45 Klein Klouwenberg PMC, Cremer OL, van Vught LA, et al. Likelihood of infection in patients with presumed sepsis at the time of intensive care unit admission: a cohort study. Crit Care 2015; 19: 319.
46 Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. Proc Mach Learn Res 2017; 68: 361–76.
47 Brunham BA, Hernán MA, Hazeue SJPA, Robins JM. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Stat Med 2004; 23: 749–67.
48 Sos HC, Lewis RJ. Pragmatic trials: practical answers to “real world” questions. JAMA 2016; 316: 1205–06.
49 Roche N, Reddell H, Martin R, et al. Quality standards for real-world research. Focus on observational database studies of comparative effectiveness. Ann Am Thorac Soc 2014; 11 (suppl 2): S99–104.
50 Ruderatina VA, Glickberg BS, Avila P, Harding-Theobald F, Wang C, Butte AJ. Accuracy of medical billing data against the electronic health record in the measurement of colorectal cancer screening rates. BMJ Open Qual 2020; 9: e000856.
51 Lauritsen SM, Thiessen B, Jørgensen MJ, et al. The framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. NPJ Digit Med 2021; 4: 158.
52 Alikhan R, Bedenis R, Cohen AT. Heparin for the prevention of venous thromboembolism in acutely ill medical patients (excluding stroke and myocardial infarction). Cochrane Database Syst Rev 2014; 5: CD001787.
53 Amin A, Neuman WR, Linghol-Smith M, Menges B, Lin J. Influence of the duration of hospital length of stay on frequency of prophylaxis and risk for venous thromboembolism among patients hospitalized for acute medical illnesses in the USA. Drug Con Stat 2019; 8: 212568.
54 Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: what data are observed and why? EGEMS (Wash DC) 2016; 4: 1203.
55 WHO. International statistical classification of diseases and related health problems (ICD). https://www.who.int/standards/classifications/classification-of-diseases (accessed Nov 19, 2021).
56 Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data 2012; 6: 15.
57 Olszewsky V, Doosti M, Vladasco C, Benecke J. Time series analysis and forecasting with automated machine learning on a national ICD-10 database. Int J Environ Res Public Health 2020; 17: 4979.
58 Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J, Doctor AI: predicting clinical events via recurrent neural networks. Proc Mach Learn Res 2016; 56: 301–18.
59 Ghasssemi M, Pimentel MA, Naumann T, et al. Multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. Proc Conf AAAI Artif Intell Intell 2015; 2015: 46–53.
60 Ellenberg JH. Selection bias in observational and experimental studies. Stat Med 1994; 13: 557–67.
61 Westreich D. Berkson’s bias, selection bias, and missing data. Epidemiology 2012; 23: 139–64.
62 Haneuse S, Schildcrout J, Crane P, Sonnen J, Breitner J, Larson E. Adjustment for selection bias in observational studies with application to the analysis of autopsy data. Neuropediatrics 2009; 32: 229–39.
63 Society of Critical Care Medicine. Using multiple imputation to avoid bias from missing data in critical care research. 2018. https://scccm.org/Communications/Critical-Connections/ (accessed Aug 2019).
64 Haneuse S, Schildcrout J, Crane P, Sonnen J, Breitner J, Larson E. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016; 183: 758–64.
65 Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016; 16: 138.
66 Haneuse S, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016; 16: 138: 758–64.
67 Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016; 16: 138.
68 Wong AK, Charpignon M, Kim H, et al. Analysis of discrepancies between pulse oximetry and arterial oxygen saturation measurements by race and ethnicity and association with organ dysfunction and mortality. JAMA Netw Open 2021; 4: e2113674.
69 Emtian M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. Chest 2020; 158: 521–28.
70 Sauer CM, Darin TA, Celi LA, et al. Systematic review and comparison of publicly available ICU data sets—a decision guide for clinicians and data scientists. Crit Care Med 2022; 50: e851–88.
71 Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in healthcare. Lancet Digit Health 2020; 2: e489–92.
72 Wegener W, Kieniewicz P. Application of imbalanced data classification quality metrics as weighting methods of the ensemble data stream classification algorithms. Entropy (Basel) 2020; 22: 849.
Copyright © 2022 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.