Research Week 2023

Unique biases in electronic health records data for research: a systematic review

Frances Hsu; Nicole Weiskopf

Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University

Keywords

Electronic health record; bias; systematic review; clinical research

Abstract

Background

Observational data fill in knowledge gaps when randomized controlled trials (RCTs) are not ethical or feasible. Electronic health records (EHRs) provide a source of observational data from diverse patient samples that reflect real-world settings. However, observational data are susceptible to biases and EHR data contain new forms of bias(es) because they were collected for clinical and not research purposes.

Methods

A systematic search for computerized medical records system and bias was conducted in Ovid MEDLINE through January 2023. English-language articles investigating bias in EHR or simulated EHR data from a country with widespread EHR adoption were included. Articles on EHR data from primary care, community clinic, or hospital were included while administrative (e.g., billing), registry, and RCT data were excluded. Biases were categorized as confounding, selection, or information/measurement bias. The source of each bias was also classified based on the EHR data-generation pathway, from clinical care to data recording, to database transformation and storage, and finally to extraction for analysis.

Results

The healthcare process introduces considerable bias into EHR data. Lack of access to healthcare reduces disadvantaged patients' data completeness, contributing to differential misclassification of confounding variable, exposure, and/or outcome. Depending on the relationship between these variables, misclassification can bias results towards (false negative) or away (false positive) from null hypotheses. Conversely, prevalent users are preferentially included in the research cohort because they are more likely to use specialty care and generate more complete data. Hence, data source (primary versus specialty care) and cohort selection criteria could introduce collider bias that distorts findings.

Conclusions

As more clinical research and machine learning algorithms utilize EHR data, researchers should carefully consider the data source, research design, data quality, and potential biases when using EHR data for research. Well-designed observational studies could improve the reproducibility and clinical value of EHR-based research.