Race and ethnicity biases introduced by filtering electronic health records for patients with "complete data"

Yasaman Fatapour[1], Jose M. Acitores Cortina[1], Nicholas P Tatonetti[1]

[1] Department of Computational Biomedicine, Cedars Sinai Medical Center, Los Angeles, CA

Background

In the current healthcare system, clinical data, particularly data collected during patient care, holds immense importance. This data plays a crucial role in prognosis, clinical trial matching, and decision-making processes, thereby enhancing patient outcomes and operational efficiency. [1]

The widespread adoption of electronic health records (EHRs) over the past decade has significantly increased the availability and accessibility of electronic clinical data. However, these available clinical data face significant challenges that hinder their effectiveness and reliability. One major issue is the lack of diversity in datapoints, particularly from underrepresented communities. [2] This underrepresentation can lead to biased research outcomes and exacerbate health disparities. Furthermore, there is a scarcity of complete longitudinal datasets, which are essential for understanding long-term health trends and treatment outcomes. [3]

One potential solution to this problem is to combine multiple datasets to reconstruct patient records from various care sites. This involves integrating EHR data with other sources, such as claims data. However, significant issues hinder the successful linkage of claims data to EHR data, including regulatory oversight, privacy concerns, data rights, and the absence of a universal patient identifier. [4] [5] Additionally, the lack of standardized data collection methods further complicates this approach, making its success questionable.[6] Addressing these challenges is crucial for leveraging big data to its full potential in advancing healthcare equity and effectiveness.

Another approach is to investigate cohort populations by including only complete data for patients, which may vary depending on the condition being studied. In this study, we aim to evaluate the effect of data completeness filters on different datasets and their impact on the patient cohort. Specifically, we examined race and ethnicity biases introduced by applying common filters to four distinct OMOP or similar structure databases, including All of Us, UK Biobank, and two geographically distinct academic medical centers. By analyzing the available data and applying each filter, we can investigate the potential biases these filters may introduce.

Methods

To examine the race/ethnicity biases introduced in the cohorts, we selected 16 commonly used filters in electronic health records research on the availability of different types of data. Filters can be grouped into three categories. The first category is based on patient demographics. This includes filters that check whether both the patient's age and sex (AgeSex) are available, and age cut off at different thresholds. The Alive filter excludes patients known to be deceased, while the zip code filter selects patients with a known zip code address. The second category is the fact type filter, which checks whether patients have at least one recorded instance of various medical data types, such as diagnoses, medications, and outpatient visit. The last category is the observation period filter, which selects patients who have had multiple interactions with the healthcare system during a specific period of time.

Then we evaluated the effect of applying these filters on self-reported race and ethnicity. This assessment was performed across all the four data sets comprising approximately 12 million patients. It is important to note that this study's purpose was to estimate the biases introduced by different types of filters that researchers might use to assess data completeness in EHR datasets. We do not claim that these are the most common filters by researchers or the optimal filters for selecting patients with complete data.

Results

Applying the observation period filter led to a substantial reduction in data availability across all races and ethnicities in all four datasets. Although there was considerable variation across healthcare systems, most patients generally had demographic data. However, among those examined, the availability of data in the white subgroup remained consistently higher compared to other racial groups after the application of each filter. Conversely, the Black/African American group was the most impacted by each filter in three datasets. Figure 1 represents the effect of each filter in data availability on Cedars-Sinai dataset. This finding highlights the potential biases that may arise in studies involving underrepresented groups**.**

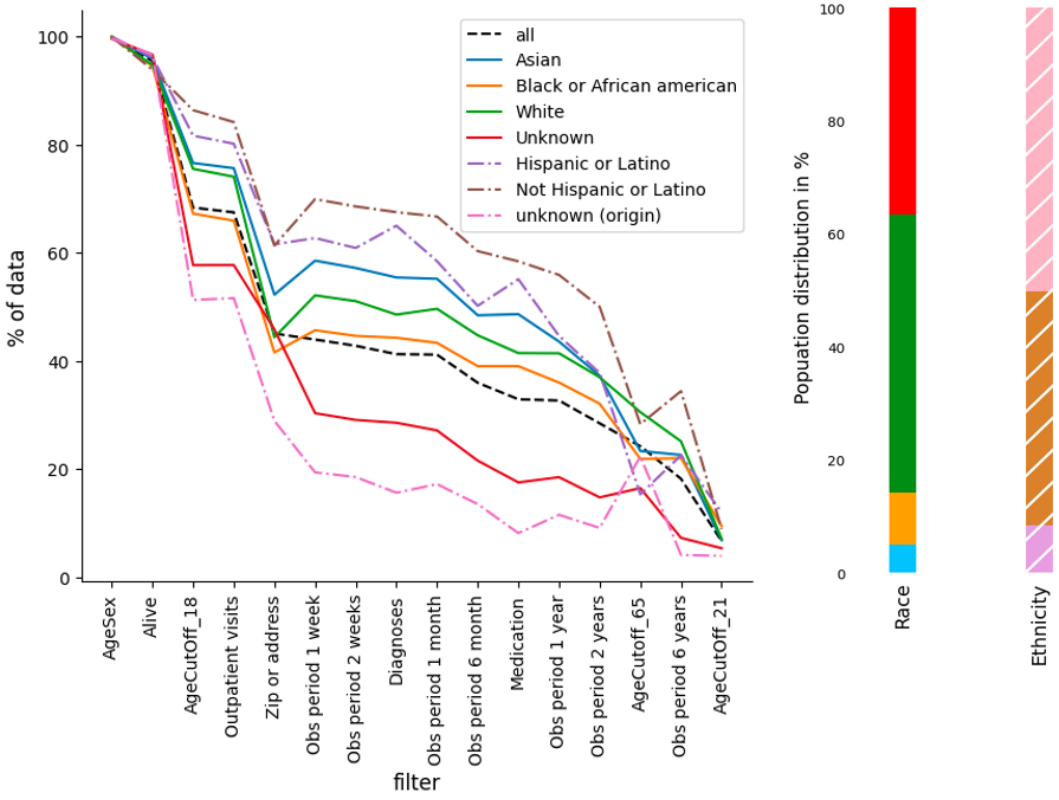

Figure 1: Data availability within Cedars-Sinai dataset after applying each filter and its impact on different races and ethnicity groups.

Conclusion

Our findings underscore the importance of using only necessary filters as they might have consequences on the diversity and completeness of population data which particularly affects underrepresented communities. This also underscores the necessity for a unified approach to data collection. Utilizing standard models, such as the OMOP common data model, can help create analogous datasets across different institutions facilitating comparison and integration. In conclusion, while more effort is needed to address the issue of missing data, researchers must also consider the unintentional biases introduced during data-driven research. Exploring techniques to mitigate the impact of these biases, such as probabilistic methods or utilizing machine learning and artificial intelligence, is crucial for improving the integrity and accuracy of their findings.

References

1. Toh C, Brody JP. Applications of machine learning in healthcare. Smart manufacturing: When artificial intelligence meets the internet of things. 2021 Jan 14;65.

2. Smith-Doerr L, Alegria SN, Sacco T. How diversity matters in the US science and engineering workforce: A critical review considering integration in teams, fields, and organizational contexts. Engaging Science, Technology, and Society. 2017 Apr 2;3:139-53.

3. Weber GM, Adams WG, Bernstam EV, Bickel JP, Fox KP, Marsolo K, Raghavan VA, Turchin A, Zhou X, Murphy SN, Mandl KD. Biases introduced by filtering electronic health records for patients with “complete data”. Journal of the American Medical Informatics Association. 2017 Nov 1;24(6):1134-41.

4. Holmgren AJ, Adler-Milstein J. Health information exchange in US hospitals: the current landscape and a path to improved information sharing. Journal of hospital medicine. 2017 Mar;12(3):193-8.

5. Heintzman, John, et al. "Agreement of Medicaid claims and electronic health records for assessing preventive care quality among adults." Journal of the American Medical Informatics Association 21.4 (2014): 720-724.

6. Adler-Milstein J, Jha AK. Sharing clinical data electronically: a critical challenge for fixing the health care system. JAMA. 2012 Apr 25;307(16):1695-6.