Analyzing Demographic Data Gaps in the PhysioNet Database: Toward Mitigating AI Bias in Biosignal Algorithms

Sarah Jiang[1,2], Perisa Ashar[1], Dr. Md Mobashir Hasan Shandhi[1], Dr. Jessilyn Dunn[1,3,4]
[1]Department of Biomedical Engineering, Duke University, [2]Department of Computer Science, Duke University,
[3]Department of Biostatistics & Bioinformatics, Duke University, [4]Duke Clinical Research Institute, Duke University

Introduction

The **PhysioNet Open Access Database**[1] is among the most extensive and utilized repositories of biosignal data worldwide, with 181 datasets that are crucial for training and validating artificial intelligence/machine learning (AI/ML) algorithms designed to estimate and predict various health-related outcomes. However, the lack of uniform demographic reporting across these studies introduces a significant risk of bias, particularly affecting underrepresented and underserved populations.

Algorithms trained on non-representative data may skew results and exacerbate existing health disparities, and cannot be generalized across real-world populations, and thus are limited in their implementation in real-world healthcare settings. We sought to identify existing relationships between demographic data that is more frequently reported, such as study participant age and sex, and demographic data that is frequently absent from biosignal studies, such as race and ethnicity.

Methods

We conducted a *systematic analysis of these 175 biosignal datasets/databases* involving human subjects (as of July 2023) to identify reporting patterns primarily regarding four key demographic variables: race, ethnicity, sex, and age, as well as supplementary information such as study size (N), date of publication, location, and biosignal type. Where detailed data, on the level of individual participants, was available, this data was used to calculate our values. We *conducted six pairwise Chi-square tests of independence* (Bonferroni α = .0083) on all permutations of two variables chosen from race, ethnicity, age, and sex to identify relationships in demographics reporting.

Results

Reporting Frequencies:

- 6.9% (N=12) of the 175 studies involving human participants reported all four demographic variables
- 14.3% (N=25) report no demographic information
- Only 13.1 % (N= 23) reported race and only 8.0% (N = 14) reported ethnicity
- 81.1% (N = 142) report sex and 79.2% (N = 139) report age

Reporting Patterns:

- Studies that reported at least two variables reported both age and sex (N = 133)
- No studies reporting < 2 variables reported race or ethnicity (N = 42) (Figure 1).
- Within studies that include ethnicity data, the only reported ethnicities are latinx/non-latinx, or similar binary categories
- Reporting of age was not independent of reporting sex, and reporting of ethnicity was not independent of reporting race, and vice-versa (Figure 2).

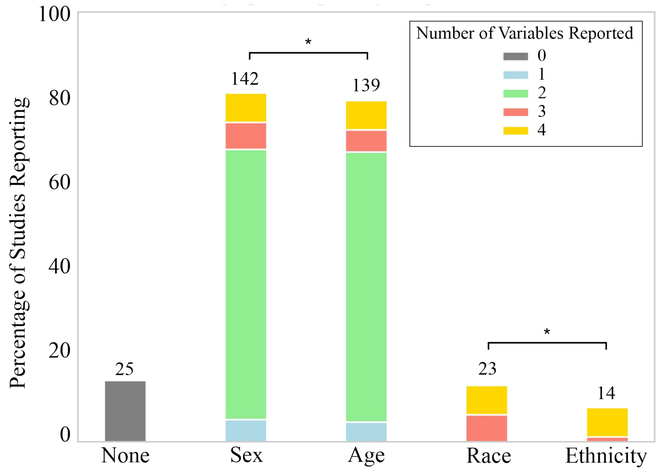

Fig. 1: Demographic Reporting Frequency. Stacked bar graph showing the reporting rates of race, ethnicity, sex, and age, as well as how many demographic variables were reported by studies. Significant associations marked by brackets and asterisks. All studies reporting race and/or ethnicity information also included both sex and age information for participants.

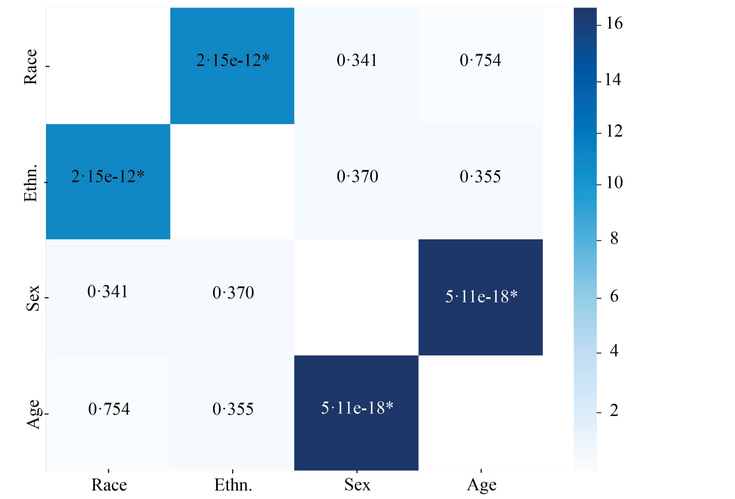

Fig. 2: Chi-Squared Tests of Independence – Heat map displaying p-values from Chi-Squared tests, with significant values marked by asterisks. P-values <0.0083 colored according to the logarithmic scale at the right, values >0.0083 are shaded in pale blue.

Conclusions

These analyses not only shed light on the prevalence of non-standardized demographic reporting within biosignal datasets, but also set the stage for evaluating the impact of such reporting on the accuracy and bias of AI/ML models in healthcare. There is an urgent need for the research community to lead in establishing and enforcing standards for comprehensive demographic reporting. By enhancing data collection protocols to ensure the representation of diverse populations, we can develop more accurate and fair algorithms to estimate and predict health-outcomes.

References

[1] "PhysioNet Databases." Accessed: Oct. 17, 2024. [Online]. Available: https://physionet.org/about/database/

Associated Publication:
S. Jiang, P. Ashar, M. M. H. Shandhi, and J. Dunn, "Demographic reporting in biosignal datasets: a comprehensive analysis of the PhysioNet open access database," The Lancet Digital Health, vol. 0, no. 0, Oct. 2024, doi: 10.1016/S2589-7500(24)00170-5.