LAY ABSTRACT

TITLE: To weight or not to weight? The effect of selection bias in 3 large electronic health record-linked biobanks and recommendations for practice

JOURNAL: Journal of the American Medical Informatics Association. 2024; 1479-1492

AUTHORS: Maxwell Salvatore, MPH[1,2], Ritoban Kundu, MS[2,3], Xu Shi, PhD[3], Christopher R. Friese, PhD[4,5,6], Seunggeun Lee, PhD[3,7], Lars G. Fritsche, PhD[2,3,4], Alison M. Mondul, PhD[1,4], David Hanauer, MD[8], Celeste Leigh Pearce, PhD[1,4], Bhramar Mukherjee, PhD[1,2,3]*

*Corresponding author

INSTITUTIONS:

1 Department of Epidemiology, University of Michigan, Ann Arbor, MI 48109-2029, United States

2 Center for Precision Health Data Science, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, United States

3 Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, United States.

4 Rogel Cancer Center, Michigan Medicine, University of Michigan, Ann Arbor, MI 48109-2029, United States.

5 Center for Improving Patient and Population Health, School of Nursing, University of Michigan, Ann Arbor, MI 48109-2029, United States.

6 Department of Health Management and Policy, University of Michigan, Ann Arbor, MI 48109-2029, United States.

7 Graduate School of Data Science, Seoul National University, Gwanak-gu, Seoul, Republic of Korea.

8 Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI 48109-2054, United States.

This is attributed to the CEECR grant: UG3CA267907

LAY ABSTRACT

Public health research seeks to understand the drivers of health and disease in a population like US adults. Sometimes, researchers analyze data characteristics different from the population they want to study. To account for these differences, researchers apply a statistical tool called weights. Researchers apply more weight to groups that need more representation in the data. The goal is to make the data look more like the population of interest.

For example, almost 1 in 7 people in the US live in a rural area, but a dataset may only have 1 in 20 people from a rural area. So, to better represent the rural population in the

data, researchers would apply more weight to data from rural areas to make the dataset look more like the US population.

Suppose researchers ignore differences between the data and those in the population. In that case, analyses may give us incorrect information about the population. However, when researchers apply weights correctly, analyses give us more accurate information about the population.

In this study, the researchers created rules to help scientists better apply weights to their data. To make these rules, the researchers tested weights on 3 large electronic health record (EHR) datasets. EHR datasets contain medical information on many people and are common to use in research on populations.

Each EHR dataset collected data differently, and groups in these datasets differed from their populations of interest. The researchers created weights so that data would represent these populations.

The researchers ran different analyses to see how the weights would change results. Some analyses had more accurate results when weights were used. For example, using weights gave more accurate results when analyzing the relationship between colorectal cancer and different diseases.

This research tells us that studies using EHR data should explain how groups are represented in their data, and how they use weights. Researchers need to think about the goal of their analyses, decide which groups to study, and use the correct weights.