

LAY ABSTRACT

TITLE: A doubly robust framework for addressing outcome-dependent selection bias in multi-cohort EHR studies

JOURNAL: Biostatistics, 2026, Vol. 27, Issue 1

AUTHORS: Ritoban Kundu^{1, *}, Xu Shi¹, Michael Kleinsasser¹, Lars G. Fritsche¹, Maxwell Salvatore², and Bhramar Mukherjee^{3,4}

* Corresponding author

INSTITUTIONS:

1 Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, United States of America.

2 Biomedical and Translational Informatics Laboratory, Perelman School of Medicine, University of Pennsylvania, United States of America.

3 Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, United States of America.

4 Department of Epidemiology (Chronic Disease), Yale School of Public Health, New Haven, Connecticut, United States of America.

This is attributed to the CEECR grant: UG3 CA267907

LAY ABSTRACT

Hospitals and clinics have Electronic Health Records (EHRs) that have patient demographic and medical data. This patient data in EHRs help researchers study and understand risk factors for diseases, such as air pollution, pesticides, or industrial chemicals that might increase the risk of developing cancer. However, EHRs only include information from people who actively seek medical care at the specific hospitals or health systems the data comes from. Patients who are very sick or undergo specialized treatment are more likely to have more detailed health records than healthy individuals. This means that when researchers use data from the EHR, they may not have data from healthy patients or all people in the population, and the studied chemical may appear to be more related to a disease than it really is.

To solve this problem, researchers from the Michigan Cancer and Research on the Environment Study (MI-CARES) team created a method called Joint Augmented Inverse Probability Weighting (JAIPW). MI-CARES' study team wants to understand the impact of known and suspected environmental exposures on the health of Michigan residents, particularly cancer risk. This new JAIPW method helps researchers use EHR data that comes from multiple hospitals and clinics. The research team used data from the Michigan Genomics Initiative (MGI) and the National Health and Nutrition Examination Survey (NHANES) to test their JAIPW method. NHANES is a national survey that measures the health and nutrition of adults and children in the United

States. The JAIPW method successfully confirmed standard cancer metrics that come from SEER – a large dataset on cancer incidence and survival in the US adult population.

The new JAIPW method can help CEECR researchers study environmental factors better and whether they contribute to cancer risk.