TITLE: Privacy-enhancing sequential learning under heterogeneous selection bias in multi-site electronic health records data

JOURNAL: Journal of the American Medical Informatics Association, 2026, 1-14

AUTHORS: Ritoban Kundu[1, *], Maxwell Salvatore[1,2], Kumar Kshitij Patel[3], Lucila Ohno-Machado[4,5], Hyunghoon Cho[4], Xu Shi[6], and Bhramar Mukherjee[7,8,9]

* Corresponding author

INSTITUTIONS:
1 Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania, Philadelphia, PA, United States of America.
2 Department of Genetics, University of Pennsylvania, Philadelphia, PA, United States of America.
3 Foundations of Data Science, Yale University, New Haven, CT, United States of America.

This is attributed to the CEECR grant: UG3 CA267907

LAY ABSTRACT

Each patient at a hospital or clinic has an Electronic Health Record (EHR) that includes information about their demographics and health. These EHRs from hospitals and clinics contain large amounts of data from many patients and can help researchers study and understand diseases. Researchers can combine EHRs from multiple hospital systems across the country to study health patterns such as how exposure to pollution and environmental chemicals might increase cancer risk in different communities.

However, combining EHRs can be difficult. Health systems have privacy laws and strict security measures to ensure individual patients' private health information is protected. So, hospitals cannot simply share all data in an EHR because it includes patient names, diagnoses, and other personal information. Researchers also must consider how the patients included in the EHRs are representative of the larger population. For example, the demographics and health of patients seen at a large, specialized cancer center are different from the patients seen at a small rural clinic. This means that the patients represented in EHRs are different. Researchers must consider these differences between patient populations so the conclusions they draw from the data are accurate and apply to larger populations.

To solve these problems, researchers created a new method to use data from multiple EHRs called privacy-enhancing sequential learning. Instead of sharing EHRs from different centers to create one large dataset, researchers send a mathematical model from hospital to hospital. Each hospital updates the model using its own EHR without ever having to share the whole EHR. This prevents sharing private patient information.

As the model moves from site to site, it automatically adjusts for each hospital's unique patient population. This new method allows scientists to safely use data from many patients across different regions without putting personal privacy at risk. By correctly accounting for differences between health systems, researchers can get faster, more accurate answers about how environmental exposures and other risk factors impact human health.