PROJECT INFORMATION

May 2021

Developing Statistical Methods for Estimating Patient Phenotypes Using Electronic Health Record Data

Principal investigator

Rebecca Hubbard, PhD

Organization

University of Pennsylvania

What was the project about?

Researchers can use data from electronic health records, or EHRs, in studies that compare two or more treatments. In these studies, researchers need to identify all patients with the same phenotype.

Phenotypes are a person's known traits, like height and weight, or known health problems, like diabetes. However, in EHR data, some data on patient traits or health problems may be missing for some patients.

Missing data in EHRs make it hard to correctly identify all patients with the same phenotype. It's even harder when data are missing due to a patient's health status. For example, patients with uncontrolled diabetes may need more lab tests than patients with controlled diabetes. As a result, researchers who are looking at lab tests may not identify patients with controlled diabetes as having diabetes.

In this project, the research team developed and tested a new statistical method that accounts for missing EHR data to estimate patient phenotypes.

What did the research team do?

The research team developed the new method. The method combined multiple types of EHR data, such as lab results, diagnostic codes, and patient age and gender to better estimate patient phenotypes. It also combined information on the patterns of missing data and the available data for each patient.

Next the research team tested the method using data created by a computer program. The data looked like

real EHR data for 1,000 patients at risk for type 2 diabetes. The team created two types of missing data: data missing by chance and data missing due to a patient's health status. Then the team looked at how well the method worked compared with four other methods that estimate patient phenotypes.

Doctors, patients, and caregivers helped design the study.

What were the results?

The new method worked better than the other methods to estimate phenotypes. It was 99 percent accurate compared with 94 percent for the next best method.

What were the limits of the project?

The project applied the method to data that met a specific set of conditions. Results may differ for data that don't meet these conditions.

Future studies can test the methods with data to detect phenotypes that change over time, such as cancer stage.

How can people use the results?

Researchers can use the new method to identify patients with specific health problems for research, even when the EHR has missing data.

To learn more about this project, visit www.pcori.org/Hubbard373.