Statistical Methods for Handling Missing Data in Large Research Studies

Principal investigator
Qi Long, PhD

Organization
Emory University

What was the research about?
Health registries record data about patients with a specific health problem. These data may include age, weight, blood pressure, health problems, medical test results, and treatments received. But data in some patient records may be missing. For example, some patients may not report their weight or all of their health problems.

Research studies can use data from health registries to learn how well treatments work. But missing data can lead to incorrect results. To address the problem, researchers often exclude patient records with missing data from their studies. But doing this can also lead to incorrect results. The fewer records that researchers use, the greater the chance for incorrect results.

Missing data also lead to another problem: it is harder for researchers to find patient traits that could affect diagnosis and treatment. For example, patients who are overweight may get heart disease. But if data are missing, it is hard for researchers to be sure that trait could affect diagnosis and treatment of heart disease.

In this study, the research team developed new statistical methods to fill in missing data in large studies. The team also developed methods to use when data are missing to help find patient traits that could affect diagnosis and treatment.

What were the results?
The research team's new methods worked well in studies with many patients where each patient had many traits. Using data from a stroke registry, the team found patient traits that predicted how long patients
- Were at the hospital before receiving a brain scan to diagnose a stroke
- Stayed in the hospital after a stroke

What did the research team do?
The research team first tested the new methods using data created to mimic data from large studies. Next, the team tested the methods using real data from the Georgia Coverdell Acute Stroke Registry. The registry has data from 86,322 stroke patients in Georgia from 2005 to 2013. The team used data from all 86,322 patients to look at how long it took patients to get a brain scan for diagnosis. They used data from 1,807 patients to look at how long patients needed to stay in the hospital.

In addition, the research team made computer programs for the new methods that other researchers could use.

During the study, the team worked with a group of advisers and researchers. The group was from the Georgia Department of Public Health, the Georgia Coverdell Acute Stroke Registry, and the Centers for Disease Control and Prevention.

What were the limits of the study?
The new methods apply only to data that are missing by chance. These methods require powerful and expensive computers. Future research could try to
develop methods that work well but don't require such expensive computers.

**How can people use the results?**
Researchers can use the new methods and computer programs to get results that are more accurate in large studies with missing data. Such studies can help patients and their doctors compare treatments.

*To learn more about this project, visit pcori.org/Long078.*