SHARE

NEW PERSPECTIVES

Choosing to Work with Large Datasets: Perspectives of a Junior Researcher

Neda Laiteerapong, MD, MS

Dr. Laiteerapong is faculty at the University of Chicago, Section of General Internal Medicine.

Large datasets provide a great opportunity to hone analytic skills using "real" data while producing publishable findings. As a junior researcher, I've had the opportunity to work with several large datasets. There are two key features to understand before deciding to work with large datasets. First, it is important to have a sincere love of data and data analysis. Second, the large dataset must be essential to answering your research question because a secondary data project is not necessarily faster or easier than a primary data project.

From my perspective, loving data is essential to working with large datasets because as a "database" researcher, data are your "participants" and statistical software is your "setting." Many researchers may find working with numbers instead of participants to be quite boring. Personally, I relish the opportunity to work with data so that I can better understand not only my research but also the world I live in. Just this week, I collected data on how long it took me to preview a clinic chart (average: 6.1 minutes) and see a patient (average: 17.3 minutes). Multiplying the excess time per patient by the number of patients seen easily explains why I often run late. When deciding if I should move closer to work, I performed a regression analysis and found that my baseline morning commute time was 23 minutes and that the afternoon commute took at least 45 minutes no matter what I did. Within a year, I moved and now live within walking distance of work.

As a researcher, data are only valuable when I am able to transform them into meaningful statistics. Large datasets tend to be generalizable to big populations, like the US civilian non-institutionalized populations represented in the National Health and Nutrition Examination Survey (NHANES); the Medical Expenditure Panel Survey (MEPS); or the National Social Life, Health, and Aging Project (NSHAP). This value is not lost on researchers, reviewers, or editors. Therefore, as I was starting my research career, I decided that it was essential that I learn to analyze a few large datasets. Another benefit of using large datasets is that many are publically available, which means that for a junior researcher with a small budget, the start-up costs are quite low.

There are, however, several major challenges to using national datasets. First, because many national datasets are publicly available, your research question may have already been answered completely or partially by another researcher. Second, since the dataset was created for a reason other than answering your research question, key variables to your research may be missing or incomplete. Third, the dataset may have been collected or organized in such a way that your research will not fit the dataset.

In response to these challenges, I have employed three strategies to produce novel publishable work. The first is to *use restricted data that are linked to publicly available data*. Since there are barriers to accessing restricted data, researchers may have not studied these additional variables, thus making the chance that your research question is novel much greater. For example, NHANES includes genetic, geographic, mortality, Medicare claims, social security benefits, and air quality data that are not publicly available. In my prior work, I was interested in studying how health center use was associated with health care utilization and quality of care. Since patients often receive care relatively close to home, it was important to adjust for the distance between each patient's home and his/her usual source of care. This type of analysis had not been done before and required restricted data from MEPS on health center and household addresses. Gaining access to the restricted data took time, effort, and planning, but because of the time I invested in the Agency for Healthcare Research and Quality (AHRQ) data center, I was able to produce new findings demonstrating that health center use was associated with lower utilization of care of the same or better quality.[1]

The second strategy is to *ask questions that require the combination of more than one dataset*. Since cross-sectional analyses are relatively easy, many will have already been done using national datasets. However, asking important but more complicated research questions that require longitudinal data leverages the benefits of panel data and increases the chances that your question will be novel. For example, MEPS is a set of nationally representative two-year panel surveys, but many researchers only include one year of MEPS data in analyses, which ignores the benefit of MEPS' longitudinal insurance data.

The last strategy is to *use lesser-known datasets*. One of the first things I did after I decided to use a

continued on page 2

continued from page 1

national dataset was to research publicly available datasets to understand the range of questions that could be answered with these sorts of data. As a result of this search, I worked with data from the NSHAP (http://www.norc.org/Research/Projects/Pages/national-social-life-health-and-aging-project.aspx), which is a lesser-known, longitudinal, population-based dataset of health and social characteristics of older community-dwelling US adults.[2]

To get started using large datasets, it is important to familiarize oneself with the datasets and what they have to offer. I have found that each dataset has unique opportunities for establishing inclusion and exclusion criteria (e.g. race/ethnicity, age groups, households, doctor visits), years of data, and sources of variables (e.g. patient self-report, laboratory, examination, visits). In order to find a research question that has not been previously studied, I try to see if the data have been linked to other datasets. For my first large dataset project, I looked for a relatively "beginner-level" dataset to get my feet wet. NHANES, for example, has a very easy-to-follow tutorial on its structure and also a listserv that new researchers can join. Lastly, as with all research projects, it is crucial to do a literature review, but the literature review for using datasets is a little different. I have found it extremely helpful to do a literature search on prior use of the dataset in order to find out how user friendly it is and to understand its lesser-known features.

Just like every other research project, large dataset research takes time, energy, and interest. As a data-driven clinical researcher, I seek out opportunities to use data in my everyday life, and so I enjoy the type of work that large datasets require. With creativity and passion for discovery, I think that large datasets provide an immense opportunity to produce meaningful research.

References

1. Laiteerapong N, Kirby J, Gao Y, Yu TC, Sharma R, Nocon R, et al. Health care utilization and receipt of preventive care for patients seen at federally funded health centers compared to other sites of primary care. Health Serv Res 2014; 49(5):1498-518.
2. Laiteerapong N, Iveniuk J, John PM, Laumann EO, Huang ES. Classification of older adults who have diabetes by comorbid conditions, United States, 2005-2006. Prev Chronic Dis 2012; 9:E100.

SGIM