User's Guide: Which Comes First - The Dataset or the Research Question?

Researchers often face a decision: whether to first develop a research question and then find the best dataset to answer it, or to first select a high-value dataset and then develop the research question. For most researchers, the best approach is often a hybrid of these two options.

The core of high-quality research is a cogent and important research question. For this reason, it is usually a mistake to simply choose a dataset and then flip through the codebooks until one finds an interesting variable. On the other hand, datasets rarely have all of the data that one wants, so trying to fit a pre-conceived notion into existing data can be an exercise in frustration. Moreover, this approach often results in subpar research when the investigator doesn't respect the limitations and strengths of the data with which she is working.

Thus, the best approach is often to develop a broadly-defined area of inquiry, and then identify a handful of datasets that are well-suited to that focus. Then, one can carefully evaluate the structure and content of the data to look for unique ways to translate that area of inquiry into a specific question that is both important and well-suited for that dataset. For example, a researcher might find that a dataset contains a unique series of questions that provide a novel framework for studying her area of interest. Or, the dataset may have a unique structure – such as the collection of longitudinal data, or national representativeness, or linking of patient survey data with biomarkers – that offers a fresh and exciting way to evaluate a research topic.

Finally, the accessibility, ease of use, and local experience working with a dataset is of critical importance. For example, Medicare claims data is a tremendous resource for research, but is very challenging to use. If one's mentor has used this database and one has access to local data analysts with extensive experience using Medicare data, that's great. If not, proceed with caution unless you have an abundance of time, money, and patience. For this reason, the best datasets for a junior investigator are often those where (1) there is local experience using the dataset that can be put to use, and/or (2) the dataset is relatively easy to access, learn, and use.