User's Guide: Opportunities, Pitfalls, and Common Mistakes

OPPORTUNITIES:

Size and scope of data: The large size and scope of many datasets permits the investigator to study research questions that cannot be answered by collecting one's own primary data.

Efficiency: Because the data is already present, studies using secondary data usually can be accomplished more quickly than studies involving primary data collection. This is often of essential importance to fellows and junior faculty, who for purposes of academic advancement need to demonstrate the ability to complete research and publish papers over a relatively short time frame.

Novel research questions: Another great benefit of secondary data can be the availability of data that one might never have independently thought to collect, but which can be used to powerful effect. For this reason, carefully reading codebooks and study supplements can often generate fresh and exciting research ideas.

PITFALLS AND COMMON MISTAKES:

Getting in over your head: Datasets vary widely in their ease of use, accessibility, and cost. A common mistake among junior investigators is to attempt to use complex datasets without sufficient support, often resulting in months of wasted time and frustration before one eventually moves along to another research project. Thus, is it critical to understand the complexity of working with a dataset at the outset and to identify the resources (for example, statistical programming, bio-statistical support, and funds) that one will need to successfully complete a project. Delving into a project and then abandoning it before a paper is written is a major waste of time for a junior investigator and her mentors.

Not knowing the data: Once an investigator decides to pursue a secondary data analysis, she should strive to know the database, measures, and subjects as much as if she had collected the data herself. She should know *everything* about how subjects got into the study and how the study handled subjects over time, and she should be thoroughly familiar with the strengths and limitations of the measures employed in the study. When it comes time to write the paper, the investigator will write the methods section largely as if she had collected the data. Saying "we have a dataset" is never a method.

Data validity and generalizability: Problems often occur when an investigator fails to consider the validity of the data that she is using. For example, an investigator who is researching

socioeconomic correlates of osteoarthritis may define patients as having the disease if they answer "yes" to the survey question: "Have you ever been told by a doctor that you have arthritis?" Clearly, this question may have limited sensitivity and specificity for identifying patients with meet formal clinical criteria for osteoarthritis. This fact does not make the question worthless nor does it make the research invalid. Rather, such issues of validity need to be accounted for in interpreting the results, and by finding creative ways to increase the content validity of one's predictors and outcomes. On a related note, up-front critical thinking about the representativeness of the dataset sample can prevent downstream problems of completing a study only to realize that it has limited generalizability and may only be of limited interest to others.

Quality of research questions: The broad scope of secondary data can sometimes tempt researchers into conducting many analyses in the hopes of coming upon something with a statistically significant association – in other words, data dredging. As with any research study, the quality of the research question and a thoughtful, conceptually-driven approach to the analytic plan is essential for conducting high-quality research.