1

An Introduction to Secondary Data Analysis

What Are Secondary Data?

In the fields of epidemiology and public health, the distinction between *primary* and *secondary* data depends on the relationship between the person or research team who collected a data set and the person who is analyzing it. This is an important concept because the same data set could be primary data in one analysis and secondary data in another. If the data set in question was collected by the researcher (or a team of which the researcher is a part) for the specific purpose or analysis under consideration, it is *primary data*. If it was collected by someone else for some other purpose, it is *secondary data*. Of course, there will always be cases in which this distinction is less clear, but it may be useful to conceptualize primary and secondary data by considering two extreme cases. In the first, which is an example of *primary data*, a research team conceives of and develops a research project, collects data designed to address specific questions posed by the project, and performs and publishes their own analyses of the data they have collected. In this case, the people involved in analyzing the data have some involvement in, or at least familiarity with, the research design and data collection process, and the data were collected to answer the questions examined in the analysis. In the second case, which is an example of *secondary data*, a researcher poses questions that are addressed through analysis of data from the Behavioral Risk Factor Surveillance System (BRFSS), a data set collected annually in the United States through cooperation of the Centers for Disease Control and Prevention and state health departments. In this case, the person performing the analysis did not participate in either the

research design or data collection process, and the data were not collected to answer specific research questions.

As an example of the same data set serving as both primary and secondary data, consider the increasingly common practice of one researcher performing an analysis of data collected by a research team with whom he or she has no connection. This type of analysis is facilitated by the ease of sharing data stored electronically and the concomitant creation of electronic data archives that allow access to secondary users; some of these archives are discussed in Chapter 7. Such analyses may serve a variety of purposes, such as addressing questions not considered in the original analysis or examining how a different analytic approach might change the conclusions reached from the first analysis. In either case, the same data set serves as *primary data* for the original research team and *secondary data* for the researcher performing the later analysis.

This book deals primarily with secondary data in the sense of data sets that can be obtained and analyzed in detail by the individual researcher. There is another type of secondary data, again not mutually exclusive with the first, meaning statistical information about some geographic region or other entity. This type of information is often useful to researchers: when you place your research project in context by describing the racial makeup or median house value in the metropolitan area where you conduct your research, the data used to compute those statistics were probably secondary data. Often these statistics are computed on data collected by the federal government, and Chapter 7 discusses several websites that were created specifically to permit easy access to these types of statistics. In addition, many of the data sets described in this book are accessible through an online interface that allows the quick computation of basic statistics, without requiring the user to download data and use a statistical program to analyze it. The availability of such interfaces has been noted in the sections pertaining to each data set.

Most of the data sets discussed in this volume contain either data collected through surveys or censuses, such as the National Health Interview Survey and the U.S. Census, or administrative records such as the medical claims records submitted to the Medicare system. There are other types of secondary data, including diaries, videorecordings, and transcripts of

interviews and focus groups: some of these are included in sources discussed in Chapter 7. Data such as interview transcripts are often analyzed using qualitative data methods rather than the quantitative techniques appropriate for most of the data sets discussed in this volume. Secondary analysis of qualitative data is a topic unto itself and is not discussed in this volume. The interested reader is referred to references such as James and Sorenson (2000) and Heaton (2004).

Advantages and Disadvantages of Secondary Data Analysis

The choice of primary or secondary data need not be an either/or question. Most researchers in epidemiology and public health will work with both types of data in the course of their careers, and many research projects incorporate both types of data. A more useful approach to this question is to focus on selecting data that are appropriate to the research question being studied and the resources available to the researcher; the latter include time, money, and personal expertise. In this spirit, we offer a summary of the major advantages and disadvantages of working with secondary, as opposed to primary, data.

The first major advantage of working with secondary data is economy: because someone else has already collected the data, the researcher does not have to devote resources to this phase of research. Even if the secondary data set must be purchased, the cost is almost certainly lower than the expense of salaries, transportation, and so forth that would be required to collect and process a similar data set from scratch. There is also a savings of time. Because the data are already collected, and frequently also cleaned and stored in electronic format, the researcher can spend the bulk of his or her time analyzing the data. There is also the influence of preference: secondary data analysis is an ideal focus for researchers who prefer to spend their working hours thinking of and testing hypotheses using existing data sets, rather than writing grants to finance the data collection process and supervising student interviewers and data entry clerks.

The second major advantage of using secondary data is the breadth of data available. Few individual researchers would have the resources to collect data from a representative sample of adults in every state in the

United States, let alone repeat this data collection process every year, but the federal government conducts numerous surveys on that scale. Data collected on a national basis are particularly important in epidemiology and public health, fields that focus primarily on the health of populations rather than of individuals. In addition, some of the data sets discussed in Chapters 2 through 7 collect data using a longitudinal design, and others are designed so certain questions are included annually or at regular intervals, allowing researchers to examine the changes in health status and health behaviors in the population over time.

The third advantage in using secondary data is that often the data collection process is informed by expertise and professionalism that may not available to smaller research projects. For instance, many of the federal health surveys discussed in this volume use a complex sample design and system of weighting that allows the researcher to compute population-based estimates of health conditions and behaviors. Although a local data collection project could conceivably use similar techniques, more often a convenience sample, whose generalizability is questionable, is used instead. To take another example, data collection for many federal data sets is often performed by staff members who specialize in that task and who may have years of experience working on a particular survey. This is in contrast to many smaller research projects, in which data are collected by students working at a part-time, temporary job.

One major disadvantage to using secondary data is inherent in its nature: because the data were not collected to answer your specific research questions, particular information that you would like to have may not have been collected. Or it may not have been collected in the geographic region you want to study, in the years you would have chosen, or on the specific population that is the focus of your interest. In any case, you can only work with the data that exist, not what you wish had been collected. A related problem is that variables may have been defined or categorized differently than you would have chosen: for instance, a data set may have collected age information in categories rather than as a continuous variable, or race may have been defined as only White/Other. A third difficulty is that data may have been collected but are not available to the secondary researcher: for instance, address and phone number information for survey respondents may have been recorded by the original

research team but will not be released to secondary researchers for confidentiality reasons. If an analysis incorporating geographic information was planned, such a restriction might make the data set unusable. For these reasons, a secondary data set should be examined carefully to confirm that it includes the necessary data, that the data are defined and coded in a manner that allows for the desired analysis, and that the researcher will be allowed to access the data required.

A second major disadvantage of using secondary data is that because the analyst did not participate in the planning and execution of the data collection process, he or she does not know exactly how it was done. More to the point, the analyst does not know how well it was done and therefore how seriously the data are affected by problems such as low response rate or respondent misunderstanding of specific survey questions. Every data collection effort has its "dirty little secrets" that may not invalidate the data but should be taken into account by the analyst. If the analyst was not present during the data collection process, he or she has to try to find this information through other means. Sometimes it is readily available; for instance, many of the federal data sets have extensive documentation of their data collection procedures, refusal rates, and other technical information available on their websites or in published reports. However, many other secondary data sets are not accompanied by this type of information, and the analyst must learn to "read between the lines" and consider what problems might have been encountered in the data collection process.

Locating Appropriate Secondary Data

There is a vast quantity of secondary data in epidemiology and public health that is available to the individual researcher. However, the sheer quantity of data available, and the fact that the data are collected and archived by many different governmental and private entities, means that the process of locating appropriate secondary data is not always straightforward. In fact, this book was written to ameliorate some of the difficulties involved. There is no single process to be followed in every case, but we offer two examples of the process of locating and analyzing secondary data to address a specific research question or problem.

This section might have been better titled "achieving a fit between your research question and the data you choose to analyze" because it is often an iterative process in which a research question is posed, potential data sets are considered, the question is refined in terms of the data available, other sources of data are considered, the question is refined again, and so on. The most typical way to use secondary data for research is to begin with a research question and seek a data set that will allow analysis of that question. An alternative method is to begin by selecting from among the available secondary data sets, and then formulating a research question that may be answered using the data chosen. Although the first method conforms more to standard beliefs about how research is done, the second approach is particularly useful in classroom instruction, and both methods can produce quality research. If the researcher begins with a question and then seeks out an appropriate data set, the following generalized sequence of procedures may be useful:

1. Define the question you want to study; for instance, "How does the experience of racism affect an individual's health?"
2. Specify the population you want to study. Are you interested in children, adults, or people of all ages? What races or ethnicities do you want to study? Do you want to analyze a national sample or one confined to a smaller area? What is the range of years you would consider (e.g., you may only be interested in data collected over the last 5 years)?
3. Specify what other variables you want to include in your analysis. In this example, you might believe that it was important to have information about the respondents' race, Hispanic ethnicity, age, gender, income, and educational level in order to include those factors in your analysis. If so, you must confirm that the data you desire are contained in the data set that you choose and that they are recorded in a manner that is useful to you. If you are interested in comparing the experiences of Hispanic Blacks and non-Hispanic Blacks, information about Hispanic ethnicity would need to be recorded in the data set independently of information about race.
4. Specify what kind of data is most appropriate for your research question: for instance, can it best be addressed through a national survey, examination of hospital claims records, or transcriptions of

interviews? Also, specify if there are any specific data collection techniques you believe are particularly appropriate or inappropriate for your question. For instance, if you do not believe people would answer questions about racism honestly in a personal interview, you would not consider any data sets collected using that technique. However, if you believe that a telephone survey would be the best way to collect this information, you might begin your search by looking at surveys that used this data collection method.

5. Create a list of data sets that include information related to your research question and examine them to see if they meet your other requirements (age range included, year of collection, etc.). This is where the interactive process begins because you may have to revise either your question or your data requirements, depending on the data that are available to you.
6. Once you have chosen your data set, examine the variables you intend to use for the analysis of problems such as missing data or out-of-range values. Also, read whatever information you can find about the data collection process, data cleaning procedures, and so on in order to evaluate whether the data quality is sufficient to meet your needs. If so, continue with the analysis; if not, either devise a way to work around it (e.g., by imputing values for the missing data) or choose another data set.

How do you generate the list mentioned in step five? By any means necessary, as the saying goes. Consider the data sets described in this book, search Medline to see what data sets other researchers have used to address your topic, search the web portals listed in Chapter 7, ask other researchers for suggestions, query relevant email lists, and so forth.

If you take the approach of beginning with a data set and crafting a research question that can be addressed using it, the process is similar, but the order of events is different. In this case, you would begin by looking at the variables contained in the data set and considering how you might combine them to create an interesting question. The process can begin with a germ of an idea, which may reflect your personal interests or a question that has arisen in your work. For instance, you might be interested in how disability affects the amount of physical

activity in which a person engages. You then need to operationalize this question so it may be tested using the variables available in the data set: how will you define disability, and how will you define physical activity? At this point, a Medline search for related articles would be in order, to see how others have addressed similar questions and whether they have done so with the data set you will be using. This step will help keep you from reinventing the wheel and will place your research in context.

Alternatively, you can begin by simply looking at the variables included in the data set to see which of them interest you. For instance, if you were planning to work with the BRFSS data from 2005, you might notice that eleven states included questions on weight control procedures. You would then look at the actual questions asked and confirm that the data were actually available. Information to answer both questions can be found on the BRFSS website (http://www.cdc.gov/brfss). This process should help you refine your focus so you can craft a research question that can be answered using BRFSS 2005 data and that would add to our understanding of public health. Because the BRFSS includes racial and ethnic data, you might decide to look at racial and ethnic differences in weight control practices. Or, taking advantage of the fact that BRFSS data are identified by state of residence, you could plan to conduct a comparison of weight control practices in different states. You could also plan a multilevel analysis that combined information about state-level characteristics from the U.S. Census (e.g., racial makeup or poverty level) with the individual-level data available in the BRFSS. When you have selected the variables you will include in your analysis, confirm that they are coded (or can be recoded by you) in a manner that will support your intended analysis and that there are no major data quality issues such as large quantities of missing data.

Questions to Ask About Any Secondary Data Set

Once you have located a secondary data set that you think is appropriate for your analysis, you need to learn as much as you can about why and how it was collected. In particular, you will want to answer the following three questions:

1. What was the original purpose for which the data were collected?
2. What kind of data is it, and when and how were the data collected?
3. What cleaning and/or recoding procedures have been applied to the data?

Sources for this information include the website of the agency or other entity responsible for collecting and/or making the data available, published reports, research articles based on the data, and personal communications with relevant individuals. For instance, many of the federal agency websites include one or more contact people who are available to answer questions about the data collected by that agency, and a Medline search will often produce citations to reports and articles discussing the procedures used to collect particular data sets.

The question of determining the original purpose of the project that produced the data is important because its influence may be present in other characteristics of the data, from the population targeted to the specific wording of questions included in a survey. Because you were not involved in planning phases for the project whose data you will analyze, you need this information in order to place the data in context. To take an extreme case, you would certainly want to know if a research project on the health effects of smoking was sponsored by a tobacco company or by a nonprofit dedicated to smoking prevention. You would also like to know if there was any particular philosophy or model of health behavior that shaped the project: for instance, was a smoking cessation program structured using the Transtheoretical Model? Knowledge of the core philosophical beliefs behind a research project can illuminate the reasons for many choices made in the planning and execution of the research and will be reflected in the end product, the data you are proposing to analyze.

It is almost impossible to know too much about the data collection process because it can influence the quality of the data in many ways, some of them not obvious. To start with, you need to know when the data were collected. A data set released in 2004 may have been collected in the first 3 months of 2004 or over a 4-year period from 2000 to 2003. Second, you want to know the process by which the data were collected: was it via telephone interviews, in-person interviews, abstraction of hospital

records, or some other technique? Third, you want to know the details of the data collection process. Questions in this regard include who actually did the data collection, how extensive was their training, and how carefully were they supervised. If the data were collected through chart review, what specific instructions were given to the reviewers? If the data were collected through a survey, what was the response rate? How many efforts were made to collect data from nonresponders? If data were collected through a telephone survey, how were numbers selected? Was there any attempt to correct for the bias introduced because households without a telephone are not a random sample of all households? The issues of survey data quality are the same whether the data set is primary or secondary. For a thorough discussion of these issues, consult a reference such as de Vaus (2002) or Bulmer, Sturgis, and Allum (2006).

The third major question in working with any secondary data set is what was done to the data after they were collected. For instance, almost all data sets include some missing data. Were these data left as missing, or were values imputed, and if so, how was the imputation done? Was any data cleaning done to remove out-of-range values, and were those cases assigned missing values or was some other procedure followed? Were certain combinations of answers considered invalid, and if so, how were they treated? A famous example of this last type of procedure was the decision in the 1990 U.S. Census to recode to the opposite gender one member of a same-gender couple who declared themselves to be married. If any recoding has been done, is it possible to restore the original values? You also need to find out if data can be weighted, and if so, for what aggregations the weighting allows the production of accurate estimates (e.g., at the national level alone or at both the national and the state levels).

Considerations Relating to Causal Inference

Causality and causal inference are complex topics that can be touched on only briefly here. Issues surrounding causal inference are discussed in greater detail by Rothman and Greenland (1998) and Phillips and Goodman (2006). One of the first rules taught in a basic statistics course is "association does not prove causation," or because A is