4

COLLECTING STUDY DATA

Lauren J. Heath and Yardlee S. Kauffman

> *"Collecting data for my chart review was tedious and time-consuming. The hardest part was motivating myself to start. However, once I took the plunge, I was reminded why I was interested in the project to begin with."*
>
> —Former PGY1 Pharmacy Resident

LEARNING OBJECTIVES

- Describe data sources, uses, and limitations.
- Identify factors contributing to the decision to use prospective versus retrospective data.
- Identify strategies to ensure that data collection is valid.
- Learn how to create a functional data collection tool and assign data codes in preparation for analysis.

INTRODUCTION

Data collection is a systematic process of gathering information for research purposes. How data are collected will depend on whether the study is *retrospective* or *prospective* as well as the study design. Several data collection strategies can be employed, including interviews, queries of electronic databases, chart reviews, questionnaires or surveys, and direct observation. No matter what type of data is collected, organizing and managing data are useful skills that will allow you to transition smoothly through the phases of a research project. This chapter reviews data sources, potential issues that may arise during data collection, and suggestions for creating a data collection tool and assigning data codes in preparation for analysis.

DATA SOURCES

There are two types of data: primary and secondary.[1]

DOI 10.37573/9781585285617.004

Uses of Primary and Secondary Data

Primary data are collected specifically for research purposes and can come from clinical observations, laboratory measurements, surveys, focus groups, interviews, or participant diaries. Primary data can be used to answer both qualitative and quantitative research questions or be collected to screen for inclusion and exclusion criteria. An advantage of using primary data is that researchers have more control over data collection. For example, in a prospective trial, a laboratory parameter, such as serum creatinine, could be collected specifically within 5–7 days after starting the study medication among all study participants. This allows for greater standardization of both the timing and type of data collected, which minimizes common problems inherent to secondary data collection. However, collecting primary data can be time consuming and expensive.

In contrast, *secondary data* are initially generated by someone other than the researcher, for non-research purposes, or they are collected for a different research project. One type of secondary data is *administrative data,* which are often used for coding and billing as part of routine medical care and stored in administrative databases.[2] Other examples of secondary data include information stored in medical records, internal electronic databases, external/commercial billing databases, patient surveillance or registry data, and government databases. One of the most common examples of secondary data are data recorded during normal clinical care, such as all of the information documented during a patient visit. This can be collected via manual medical record reviews using a standardized data abstraction form. Often, the finalized dataset will be composed of secondary data arising from different sources (e.g., administrative data queries to identify potential patients then supplemented with data collected through manual chart reviews). Secondary data are rich sources of information that an individual researcher can collect efficiently. In addition, using data collected because of routine medical care provides the opportunity to study "real-world" practices. Further, use of secondary data facilitates efficient collection of longitudinal data, which allow for assessment of changes in variables over time if the desired variables have been accurately recorded.

However, there are many potential disadvantages with secondary data.[2] For example, a researcher does not have control over how or when data are collected, which can limit data validity and highlights the importance of consulting with someone familiar with the secondary data source. Practically, this can also mean that some data will be missing or not collected in the desired timeframe. This common problem is discussed later in the chapter.

The decision to collect primary or secondary data ultimately depends on the research question, whether your study design is retrospective or prospective, and the patient population being studied. Feasibility considerations, including both the timeline and budget for the research project, are also important.

Considerations for Using Secondary Data from Administrative Databases

Using secondary data is common for observational studies.[3,4] Ensuring that data have sufficient quality is necessary if an existing database is used for your research project. Assessing the database for completeness, generalizability, reliability, and validity will help you determine the data's potential strengths and weaknesses, help you draw appropriate conclusions, and identify limitations of your source data. It is important to consider whether the secondary data source has the necessary variables for your study question. Missing data are common either because they are not reported or accessible in the administrative database, or because they were never collected (e.g., height and weight missing for some patients). Details necessary for determining a medication's intended indication may also be missing. In some cases, this can be addressed by supplementing administrative data with manual chart review; however, sometimes the rationale for starting or changing a medication is not clearly documented. Prescription refill information may be accessible through pharmacy