Use of Spreadsheets for Research Data Collection and Preparation: A Primer

Krishna Juluru, MD, John Eng, MD

Successful research results from the combination of multiple elements, including an appropriate research question, study design, research method, statistical analysis, and interpretation of results. One element of research that is easy to overlook is proper data collection and preparation for analysis. If data collection or preparation is inadequately planned or executed, the data may not be analyzable by a statistician without significant effort spent on data cleaning. Even worse, the data may contain problems that can be resolved only through time-consuming revision or repeat data collection. In this review, we present some practical guidelines and best practices for preparing data that can reduce the work of subsequent analysis.

Key Words: Spreadsheets; Statistical analysis; Research methodology; PHI; HIPAA.

INTRODUCTION

Conducting successful research involves the formulation of a clear research question, selection of an appropriate study design, selection of proper data collection methods and tools, data collection, data preparation, data analysis, and, finally, thoughtful interpretation of the results (1). Many books, courses, and articles offer information on study design and statistical analysis (2,3), but typically these resources provide little advice about data collection and preparation, which is the actual work that occurs between design and analysis. Without adequate planning for data collection and preparation, a project's data can easily become overwhelming to the point of obscuring the path to proper analysis. Even worse, the data may develop problems that require time-consuming revision or repeat collection.

Clinical research projects of even minimal complexity can still produce data that require statistical analysis beyond calculating simple means, standard deviations, and *t* tests. More advanced analysis usually requires involvement of a statistician and dedicated statistical software, such as R (R Foundation for Statistical Computing, Vienna, Austria), Stata (StataCorp, College Station, Texas), SAS (SAS Institute, Cary, North Carolina), and SPSS (IBM Corporation, Armonk, New York). It is most efficient to present data in a form that is directly analyzable by the statistician and an appropriate statistical software package. In times of limited availability, it may be difficult to find a statistician with time to correct a data set's formatting problems in addition to performing the actual analysis.

In this article, we present some best practices for data collection and preparation by researchers. These practices and pieces of advice are derived from our personal experience in advising many colleagues on statistical analysis for their research projects. In our experience, we have noted common problems and difficulties associated with data collection and preparation. In some cases, these problems have required costly additional time and/or resources to resolve. Fortunately, such problems can usually be prevented by observing some relatively simple practices. We focus on primary data collection by the researcher, but the techniques can be applied to secondary data analysis where the data were collected by someone else or extracted from another source such as an electronic medical record system.

We also draw from what we have learned from teaching several refresher courses on this topic at the annual meeting of the Radiological Society of North America. Since spreadsheets are the most common form in which data are presented to a statistician, we will present some valuable tips for their use. We will draw examples from one particular spreadsheet application, Excel (Microsoft Corporation, Redmond, Washington) because it is the most widely used application of its type and is nearly identical across multiple versions of both the Windows and Macintosh operating systems. However, most of our suggestions are applicable to any spreadsheet program, including open-source applications such as OpenOffice and Google Sheets. Although there are countless books and web

Acad Radiol 2015; ■:■■–■■

From the Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, USA (K.J.); Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, Baltimore, MD, USA (J.E.). Received May 19, 2015; Received in revised form August 27, 2015; accepted August 31, 2015. Presented, in part, in refresher courses at the annual meeting of the Radiological Society of North America from 2007 to 2012. **Address correspondence to:** K.J. e-mail: juluruk@mskcc.org

http://dx.doi.org/10.1016/j.acra.2015.08.024

resources detailing every feature of Excel, we will concentrate on a few specific essential features that make data collection easier for the clinical radiology researcher.

PREPARING TO COLLECT DATA

Because of intellectual excitement or time pressures, research projects are often started by delving directly into data collection. But before making the first spreadsheet, there are several important steps that should be taken to avoid unnecessary work and subsequent delays. The first and arguably most important step in data collection is formulating the research question. All research involves answering one or more questions, and these questions should be stated explicitly and prospectively (4–6). Defining the research question first is the best way to make sure all pertinent data are collected and that time is not wasted by collecting unnecessary information. One way to ensure that all pertinent data are collected is to identify all the important parameters that may affect the intervention or outcome being studied. Sources of clinical data are often difficult to access even once, so having to go back to collect a missed data element may be impossible. Modern computing technology places no practical limit on the number of data elements that can be collected, so relevance to the research question, not simply availability, should be the determinant of whether a data element should be collected. One must also remember that a statistically significant result still has a small probability of occurring by chance alone. A clear research question and purposeful data collection help to reduce the problem of a statistically significant result being interpreted as occurring only by chance.

An equally important but often overlooked consideration at the beginning of a project is the anticipated statistical analysis. Statistical analysis is frequently considered only after the data collection has been completed. Instead, it is best to plan the statistical analysis a priori rather than post hoc. Early consultation with a statistician is often beneficial in focusing the study so it will achieve its intended goals. In additional to clearly defining the research questions, a good starting point for communication with a statistician is the creation of a data dictionary. A data dictionary is a document that defines all the variables to be gathered in a study, with a full explanation of what the variables mean and all of their possible values. Variables should be defined to contain discrete pieces of information and not represent compound information. It may be tempting, for instance, to combine a subject identifier (ID) with a date of computed tomography (CT) scan into a single field. Instead, a better practice is to define and populate separate subject ID and scan date fields.

Data types can be continuous or categorical. A defining property of continuous data values is that they share a mathematical relationship with one another. For example, tumor volume is a continuous variable because when comparing two tumors of volumes 4 cm^3 and 2 cm^3, respectively, it can be said that the first tumor is twice the size of the second. Categorical data values are usually in the form of text, but they can also be coded numerically. When coded numerically, categorical variables do not share a mathematical relationship with one another. For example, a categorical variable for gender can take the values of "male" or "female," but it cannot be said that "male" is twice that of "female." Even when variables take a numerical form, they are often still categorical. Stage IV breast cancer, for example, is certainly worse than stage II breast cancer, but the former cannot be said to be twice as worse as the latter. Cancer stage, which often takes a numerical form, would therefore be a categorical variable. Development of a data dictionary may be an iterative process between the researcher and the statistician, and this process is part of a collaborative research relationship.

A good statistical plan will identify the main outcome (dependent) and predictor (independent) variables and define what constitutes a unit of observation. Proper statistical analysis may require a data element or encoding that may not be anticipated clinically. For example, suppose a mass is recorded as "small" if it measures less than 1 cm and "large" if it measures 1 cm or more. It would be important to know if the subsequent statistical analysis will require only the dichotomized classification ("small" or "large") or whether the numerical measurement should also be captured. After completing data collection, it would be unfortunate to discover that the numerical measurement was necessary but not recorded.

Once the main outcome and predictor variables are determined, a statistician can estimate a sample size that is appropriate for answering the research question. Sample size calculations are essential in ensuring adequate time and budget to complete the study.

To provide concrete examples of the points made in this paper, we will consider a hypothetical study, the "Magic Drug Study," which uses imaging measurements to evaluate a drug intended to treat a type of lymphoma. The research question is, briefly, "Is Magic Drug more effective than standard-of-care therapy?" The outcome in this study is the effect of treatment on tumor size as measured on CT. The study design is a prospective randomized controlled trial in which subjects presenting with the particular type of lymphoma are divided into two groups: the "drug" group that will receive Magic Drug and the "control" group that will receive standard-of-care therapy. The hypothesis is that the experimental Magic Drug will cause a greater reduction in tumor size than standard-of-care therapy. Radiologists will measure the size of an index tumor on CT scans at baseline and at multiple time points during treatment. Average reduction in tumor size between the two groups will be compared at various time points, say following 6 and 12 months of treatment. We will need to collect the following data elements for each time point: subject ID, treatment group, tumor ID (in the event that a subject has more than one tumor), date of CT scan, and, of course, the measured value itself. The subject ID should be a coded identifier unique to each subject in the study and not the patient's medical record number or other form of protected health information (7). A variety of tools can be used to collect these data elements.