Data Management 101: How to Construct and Maintain a Usable Dataset

R. Curtis Bay, PhD

Introduction

We advance our understanding of the human condition by asking questions. In dentistry, these questions are best answered through formulation of hypotheses that allow us to test the validity (truthiness) of one possible answer against others. Most simply, "This new treatment is better than what we have always used," or it is not.

Clinical questions arise naturally in the practice of clinical dentistry. Frequently, they are based on the desire to use the best available practices and procedures to optimize care for patients. Answers to clinical questions are readily available in the numerous dental journals and online content that have proliferated over the past few decades using an evidence-based approach to dentistry. Much of the evidence is trustworthy. Much of it is not. The best and most trustworthy evidence is investigator-initiated, that is, arising from clinical practice and initiated by those who seek a truthful answer, untainted by financial interest. Of course, trustworthy research is the product of sound scientific methodology. Fundamental to sound methodology is the construction of a consistent and replicable plan for data acquisition, recordation and analysis.

This session focused on the basic requirements for designing, constructing and maintaining a dataset collected in the course of conducting a research study. The nature of data was also discussed, and how it serves the purpose of research, including the various types or "qualities" of data that may be collected. Some types of data (interval and ratio-level) are more informative than others (ordinal and nominal data). It is almost always best to collect the most informative type of data that can practicably be collected. Data can always be "dumbed down" by recoding, but it is very hard to "smarten-up" data once it has been collected.

Statisticians tend to like numbers and information that comes in the form of numbers. Statistical software programs are designed to analyze numbers. Methods to codify information were shared to make datasets more amenable to statistical analysis. Examples included Male as "1"; Female as "2"; Amalgam as "1"; Composite as "2"; Glass Ionomer as "3."

Very importantly, the discussion included strategies about how best to communicate with the project statistician. Researchers should initiate communication with a statistician before and after data collection forms are designed; before these forms are used; after data entry has started and before it is completed; during the statistical analysis, and after it is finished. An open line of communication with the statistician will help to ease frustration and avoid headaches for all parties involved in the process.

Along with the data, the researcher should present the statistician with a "data map," or "dictionary" indicating explicitly what each variable is, the scale on which it was collected, and what the data elements mean. Specifically, what does a "1" mean in an Excel column labeled "Gender"? A "3" which is intended to represent an ordered category (3 out of 5 on a preference scale) will be treated very differently from a "3" reflecting a nominal category (eg. Glass Ionomer). The researcher should formally document the meaning associated with each number in a written form: Word and Excel work well. It is poor form to hand a statistician handwritten notes with multiple deletions and corrections, or to convey this information orally. Doing so may result in forgotten or lost communication, and the potential downgrade in priority of the project.

Find out early on if the analysis planned for the dataset requires a "wide" or "long" format. These are very different, and converting one to the other may be tedious. Simple, one-obser-

vation-per-subject datasets are straightforward: One line per subject, column headings in the first row. If the analyst is planning a mixed-effects treatment of the data, repeated measures on each subject are typically best treated in a one-observation-per-row format, with a unique identifier for each subject repeated across rows (a "long" format). However, some analyses (e.g., repeated measures ANOVA) require that all information, across all observations for a single subject, be entered in one row: a "wide" format.

In a "long" dataset, one or more variables must be included indicating how the multiple rows for one subject differ from one another. If row 1 is for a baseline visit, row 2 is the first follow-up, row 3, end-of-trial, then a column must be created to convey this information. It might be labeled "Visit." This information, of course, must be included in the data map.

Each cell in a spreadsheet can include only one piece of information. If the subject indicates that he is White, African American and Hispanic, this requires three columns in the spreadsheet. The statistical software, on import of the spreadsheet, will interpret a cell entry of "1 2 5" as text, rather than a series of numbers. If a subject is asked to list the years in which he has had restorative dental work performed, and he lists five years, this requires five columns in the spreadsheet. Worst are the "check all that apply" formatted questions. A separate column must be included for every possible response. An endorsement of a category equals "1"; a non-endorsement should be coded as "0".

Missing data should be explicitly coded as such; not with the word "missing," but with a numeric value that could not possibly be valid for a given variable. As an example, "99" entered as a value for a Likert-type variable scored 1 to 7 is an invalid entry, and must be flagged as "missing" by the analyst. Once "99" is defined as "missing," the statistical software will ignore that particular observation in subsequent analyses. Missing values should appear in the data map so that the analyst can define them as such before beginning the analysis. Again, do not type "missing" into a column which is defined as a numeric field. The data will be imported as text, rather than numeric, and will require conversion before the analysis proceeds.

Having analyzed data for over 2000 projects during 12 years at an academic medical center, and another 10 years at a dental, medical and ancillary health sciences university, I issue this plea: CHECK AND CLEAN YOUR DATA BEFORE GIVING IT TO YOUR ANALYST!

I have re-run hundreds of analyses because the researcher failed to check his data before giving it to me. The analysis is completed; the output is sent to the researcher; we meet to go over the results. "Whoops! Those should be "7's," not "6's". Or, "Those values aren't possible for that variable." "Sorry, I should have checked my data more carefully. Would you mind re-running all of these analyses after fixing my mistakes?" Ask your analyst to run a set of descriptive statistics on the dataset, including means, standard deviations, frequencies, minima and maxima so that the numbers can be reviewed before the actual analysis begins.

And, as an aside, in spite of the fact that the popular press insists upon making "data" singular, as in "The data shows that ..." the word "data" is not singular, but plural. The singular form is "datum." When communicating with a statistician, nothing will mark you as unsophisticated as readily as asking him or her "what the data shows." Asking what the data "show" will immediately convey that you are "adept" with numbers, which will gain the statistician's respect and admiration.

In addition to a discussion about the fundamentals of data preparation such as those above, advantages and disadvantages of using databases rather than spreadsheets to capture research data were explored. Database software offers the potential for more security than software conventionally used for spreadsheets, and is highly customizable. It also requires considerably more skill to navigate, especially during the setup phase. In the case of complex datasets, with one to many relationships and/or highly sensitive content, databases may be worth the extra effort.

The session included a discussion of internet-based data collection systems, such as SurveyMonkey®, Qualtrics and REDCap™ (Research Electronic Data Capture), including the highlights and lowlights of each. Finally, a quick overview of Microsoft® Excel (spreadsheet) SPSS (statistical software) and Microsoft® Access (database) was provided, with a demonstration of how each may be used for research data collection and analysis.