Data Management

Editors' note: This article is the twelfth in a multipart series designed to improve the knowledge base of readers, particularly novices, in the area of clinical research. A better understanding of these principles should help in reading and understanding the application of published studies. It should also help those involved in beginning their own research projects.

The most important part of any research project is the planning process. The more complete the planning, the smoother the project generally runs. This is particularly true for data collection and management. The plan for collecting, managing, and storing your data should be planned before the study begins. The investigator should have a specific plan in advance for how all data will be collected, formatted, and coded (if appropriate for some data elements) and the program(s) to be used to store and ultimately to analyze the data. The purpose of this issue in the Basics of Research series is to discuss issues relevant to data management decisions.

Data Collection

The general methods for data collection will be specified in the research proposal. The investigator will identify in the proposal whether physiologic measurements, formal surveys or questionnaires, observational data, or some other approach will be used to collect the study data. Once the investigator understands the source of their data, she or he must determine how they will document it and then enter it into the computer program to be used for analysis.

The most common method of data documentation is to record the data onto an investigator-designed form. This method has the advantage of being portable to the site of data collection. The forms can be duplicated and distributed to as many data collection sites as needed. Forms can be returned easily to the principle investigator, using mail if needed for sites at a distance. Investigator-created paper forms have great flexibility and can facilitate ease of data documentation. The major disadvantage of this method is the need to transcribe the data into the computer program for data analysis. The process of transcription, like any step in data handling, may introduce errors into the data. The advantage is that there is always a paper "source document" that can be referenced if there are later questions about the accuracy of the computer data.

The most common method of data transcription is to have a staff member read and then type the data into the computer. Data are commonly entered directly into the statistical analysis program such as Statistical Analysis Software (SAS) or Statistical Package for the Social Sciences (SPSS). Some investigators prefer to enter the data first into a spreadsheet and then into the statistical program. This second method has the advantage of not requiring the data entry person to own a copy of a potentially expensive statistical program. In some cases, a spreadsheet itself may be able to do the statistical analysis. Microsoft Excel has a toolbox of available data analysis tools, but its use may require custom loading of the software. It includes correlations, chi-square, confidence intervals, and *t*-tests, among others.

Data also can be collected on forms that can be scanned directly into the computer. "Bubble sheets" are standard forms with a number for each question and a set of circles to fill in to indicate the answer. The disadvantage of this method is that the data collection form does not contain the questions. The questions must be read from a second form. This form of data collection is also limited to categorical data and cannot be used for qualitative studies.

A second option in computer-readable forms is use of an optical character recognition program such as Teleforms. These forms resemble a standard data collection form and can capture categorical, numerical, and textual data. The major disadvantages to this approach are the time needed to create the original form and the cost of the computer program to create and read the forms.

Research data also may be collected directly into the computer. An increasingly popular method is to create a web page for data entry. This method can be made user friendly, wherein the screen resembles the familiar paper form. In addition, the individual entering data can see that the data being entered match the variable name on the screen. This may decrease errors.

The major disadvantage of this method is the need to use a computer programmer who can create the web page and the database needed to store the data. With this method, the investigator must have a computer and an internet connection at the data collection site that is available at all study entry times. This approach can add considerable expense to the study if data collection is occurring simultaneously at several sites. Additionally, some people find that data entry on a computer takes longer because of the need

to be able to type and the necessity of using the tab, enter, or arrow keys after each data element.

A final, though now less common, method for data entry is entering data directly into a flat text file. This type of data entry is demonstrated in Figure 1. The main advantage of flat files is the speed of data entry. A competent typist can enter data quite quickly because there are few if any characters to type between variables. Spaces are left occasionally to allow the eye to follow columns of values, but the spaces are much fewer than needed to enter data into a spreadsheet. Once the data are entered into this flat file they can be imported into a spreadsheet or statistical program. A disadvantage of this method of data entry is that it is easier to make a mistake and harder to catch it. Numbers, once entered, are not associated directly with a variable name. Consequently, finding errors in data entry may be more difficult.

Figure 1. Flat Text File Data Entry

```
101120195 3455 11122012
102120395 3545 11021121
103120695 4443 01011022
104121095 3343 10112222
```

Data Management

Data Coding

Most statistical programs work primarily with numerical data, rather than with text data. Data coding is the process of assigning a numerical value to a qualitative response. For example, an investigator may ask for the professional background of a transport team member. The responses may be physician, nurse, paramedic, respiratory therapist, pilot, or other. To facilitate the statistical analysis or to save space in the data file, the investigator may assign a numerical value to each possible response, for example: all physicians would be coded as a 0, all nurses 1, all paramedics 2, all respiratory therapists 3, all pilots would be 4, and all others 5.

Data transformation, a variation of data coding, is the process of changing the numerical representation of a quantitative value to another value. Data can be changed to reflect a new measurement scale or to make alterations in the distribution of the data. For example, data on neonate weight may be obtained in grams. However, if older subjects also are included in the dataset, the neonate weights may need to be transformed into kilograms to maintain consistency across all subjects. In the case of family income level, there is usually an uneven distribution of subjects at the low end of the scale. Economists often use a logarithmic transformation so that the data are more evenly distributed. If the income data are reduced to their logarithmic value, the high incomes are brought closer to the lower end of the scale and provide a distribution closer to a normal curve.

Another type of data transformation is reverse scoring. When using a survey instrument, the investigator or survey developer may elect to work some responses in a positive manner and others in a negative manner. For example, one question might be to ask your level of agreement with the statement "The transport was conducted in a safe manner" and a second question your agreement with "Communication between team members was unclear." For the first question, a score indicating high agreement would be positive, but for the second question a score indicating high agreement would be negative. To maintain consistency in the meaning of a response, the second question might be reverse scored. In this case, a 5 (very high agreement) would be transformed to a 1 (very low agreement), a 4 (high agreement) to a 2 (low agreement), and so forth.

With paper data collection forms, physically coding the data can be accomplished by one of several methods. Some researchers like to put the appropriate code on the data collection form before distribution so that the code can be read from the form at the time of data entry. Others prefer to evaluate each data collection form after the data are recorded but before data entry and place the appropriate code next to the response on the form. A final way to code data is to have a separate code book and to use this as a reference during data entry into the computer. The first method requires two passes through the data collection forms. Consequently, one of the latter two approaches often is preferred. If the investigator does not wish the subject to be aware of the coding scheme or does not have room for it on the data collection form, the last option may be the best.

Although numbers are easier to analyze, the investigator must be careful that he or she does not attribute ordinal or ratio level characteristics to data that were originally nominal in nature. For example, a correlation requires that the level of measurement for both of the variables be at the interval or ratio (ie, continuous, not categorical type data) level of measurement. Coding profession as a numerical value does not then allow the investigator to do correlations between profession and another variable. The data remain at the categorical level and thus are inappropriate for inclusion in a correlation analysis.

Data Storage

Data, once collected, must be stored carefully to prevent damage or violations of subject privacy. Most institutional review boards (IRB) will require that the investigator have a plan for security of the data before data collection begins. Such plans must cover both paper forms and computerized subject data.

A commonly used method for storing paper forms containing subject data is file folders. A filing scheme that is based on subjects, location, treatment group, etc. should be identified before starting data collection. If a large number of forms will be collected for each subject, a single folder per subject is recommended. If each subject has a single form, an alternative system that logically groups data will be appropriate. Colored folders often may be used to advantage to group like data, such as time frames.

Once the filing system is set up, data forms can be immediately placed into the designated folder once they are returned. Secure storage, using a double-lock system, is highly recommended. A common method for doing so is to place data into a locked file cabinet stored in a locked office or storage room.

Security of computerized data is also required. Research computers and specific study files should all be password protected, if not fully encrypted. Access should be limited to study personnel. This also applies if the files are stored on a network drive. All files, paper or computerized, should be stored for at least 3 years after publication or at least 7 years if relating to a new therapeutic agent.

Data backup is an important part of a data storage plan. The data file should be saved at frequent intervals during data entry to prevent loss caused by power failure or other mechanical difficulty. A good habit is to save the file using "control S" or the equivalent at least every 10 minutes during data entry. The more frequent the save, the fewer data that will need to be re-entered in case of a computer malfunction. More importantly, a plan for formal backup at the end of each day is also needed. If the data are stored on a network drive, most organizations have a nightly backup process.

If the research team is saving their data to their own computer hard drive, they will be responsible for the backup system. Saving data to a CD, thumb drive, or disk (zip or separate hard drive) at the end of each day is one approach. Any such files or disks need to be marked carefully so that the most recent disk is used for the next data entry session; otherwise, data can be lost.

The best backup plan will provide for storage of backup media at a site remote from the location of the original data. This protects against major loss of data in case of fire or other disaster that might destroy an individual computer or building.

Privacy protection of the backups is as important as security of the original data. If the backup media is to be moved out of the locked office, encryption is recommended. Encryption will limit the chance of release of confidential data should the files be stolen or lost.

Data Cleaning

Every step in data handling, including computer data entry, invariably results in mistakes. Data cleaning is the process of trying to find errors and to correct them before data analysis. Several techniques can be used for data cleaning.

Programs are available that allow the investigator to enter the data twice. The computer then checks for inconsistencies between the two sets of data and notifies the individual that a mistake has been made. The primary disadvantage of this method is the increased time needed for data entry. Another disadvantage is that is possible for the same mistake to be made twice.

Another method for cleaning data as it is entered is by using controls in the data entry program. Value checking rules can be placed within some types of data entry forms. Once a rule is created, the computer will check all data entered into that field to make sure that it meets the established criteria. For example, only numbers in a select range are acceptable to be entered into a given data cell.

The method of data cleaning that catches most errors is the process of reading the data back. With this method, one investigator reads the data elements from the data entry form and the second checks the values entered by reading a copy of the data file. Inconsistencies between the source data and the computer data become obvious. A disadvantage is the labor involved in this technique.

A statistical method for data checking is to obtain a frequency distribution for all variables. The investigator can check for obvious errors by comparing the data obtained with the data that should have been obtained. For example, if gender is coded as 0 for males and 1 for females, no 2s or 3s should appear for the value of gender. If the frequency distribution shows an erroneous value, the original data collection form can be retrieved and the data re-entered.

Summary

Proper data management techniques are essential to the conduct of a scientifically sound research project. Before the study is begun, decisions should be made regarding how the data are to be recorded, source documents stored, data coded, computer entry performed, errors corrected, computer files secured, and backups performed.

Cheryl Bagley Thompson, PhD, RN, is an associate professor and assistant dean of informatics & learning technologies at the University of Nebraska Medical Center College of Nursing in Omaha. She can be reached at cbthompson@unmc.edu. *Edward A. Panacek, MD, MPH, is professor of emergency medicine and clinical toxicology at the UC Davis Medical Center in Sacramento, California.*

1067-991X/$34.00

doi:10.1016/j.amj.2008.05.001