12: Data Management

Introduction

Data management includes all aspects of data planning, handling, analysis, documentation and storage, and takes place during all stages of a study. The objective is to create a reliable data base containing high quality data. Data management is a too often neglected part of study design,[1] and includes:

- Planning the data needs of the study
- Data collection
- Data entry
- Data validation and checking
- Data manipulation
- Data files backup
- Data documentation

Each of these processes requires thought and time; each requires painstaking attention to detail.

The main element of data management are database files. **Database files** contain text, numerical, images, and other data in machine readable form. Such files should be viewed as part of a **database management systems (DBMs)** which allows for a broad range of data functions, including data entry, checking, updating, documentation, and analysis.

Data Management Software

Many DBMSs are available for personal computers. Options include:

- Spreadsheet (e.g., Excel, SPSS datasheet)
- Commercial database program (e.g., Oracle, Access)
- Specialty data entry program (e.g., SPSS Data Entry Builder, EpiData)

Spreadsheet are to be avoided for all but the smallest data systems since they are unreliable and easily corruped (e.g., easy to type over, lose track of records, duplicate data, mis-enter data, and so on.). Commercially available database programs are expensive, tend to be large and slow, and often lack controlled data-entry facilities. Specialty data entry programs are ideal for data entry and storage. We use **EpiData** for this purpose because it is fast, reliable, allows for controlled data-entry, and is open-source. Use of EpiData is introduced in the accompanying lab.

Data Entry and Validation

Data processing errors are errors that occur after data have been collected.[2] Examples of data processing errors include:

[1] Bennett, S., Myatt, M., Jolley, D., & Radalowicz, A. (2001). *Data Management for Surveys and Trials. A Practical Primer Using EpiData*. The EpiData Documentation Project. Available: www.epidata.dk/downloads/dmepidata.pdf.

[2] This is distinct from **measurement errors**, which are differences between the true state of affairs and what appears on the data collection form.

- Transpositions (e.g., 19 becomes 91 during data entry)
- Copying errors (e.g., 0 (zero) becomes O during data entry)
- Coding errors (e.g., a racial group gets improperly coded because of changes in the coding scheme)
- Routing errors (e.g., the interviewer asks the wrong question or asks questions in the wrong order)
- Consistency errors (contradictory responses, such as the reporting of a hysterectomy after the respondent has identified himself as a male)
- Range errors (responses outside of the range of plausible answers, such as a reported age of 290)

To prevent such errors, you must identify the stage at which they occur and correct the problem. **Methods to prevent data entry errors** include:

- **Manual checks** during data collection (e.g., checks for completeness, handwriting legibility)
- **Range and consistency checking during data entry** (e.g., preventing impossible results, such as ages greater than 110)
- **Double entry and validation** following data entry
- **Data analysis screening for outliers** during data analysis

EpiData provides a range and consistency checking program and allows for double entry and validation, as demonstrated in the accompanying lab.

Data Backup and Storage

A well-known computing saying goes:

> There are two kinds of computer users. Those that have lost a major chunk of data, and those who are going to lose a major chunk of data.

Data loss can be due to natural disasters, theft, human error, and computer failure. You’ve worked to hard to collect and enter data, and you must now take care of it.

The most common loss of data among students is due to “loss” of data somewhere on the computer. The best way to prevent such loss is to know the physical location of you data (local drive, removable media, network) and to use logical file names. All too often students save files to unknown locations (usually the default set up by the program) but never find saved files or have the saved files deleted by the local area network as a part of routine data cleanup. ALWAYS BE AWARE OF THE LOCATION AND PATH (“folder”) TO WHICH FILES ARE BEING WRITTEN.

In addition, it is essential to back-up all data (e.g., data files, code books, software settings, computer programs, word processing documents). Backup systems entail manual or automated copying of files to removable media (e.g., floppy disks, Zip disks, tape) or to network storage. Backup procedures should be thoroughly tested to ensure archived files remain uncorrupted and can be restored. Procedures should be written up so that personnel unfamiliar with backup and restore methods could follow them from scratch.

Health researchers must be aware of *confidentiality and ethical requirements* when working with research files. This is especially important when data contain personal identifiers and medical information. It is each researcher's duty to make him or herself aware of local, national, and international laws governing use of health data. Many legal problems can be avoided by using anonymous data files (i.e., data containing information about individuals but without personal identifiers). However, it is not always clear when in fact data become fully anonymous. For example, in studying a rare disease in an identified population, it is conceivable that an unscrupulous user could use supplementary information to re-identify individuals. Although the objective of protecting individual identity in such instances is clear, it is not always clear how far the analysts responsibility extends in protecting personal identities under given circumstances.