16. Data management and data analysis*

Data management: Strategies and issues in collecting, processing, documenting, and summarizing data for an epidemiologic study.

1. Data Management

1.1 Introduction to Data Management

Data management falls under the rubric of project management. Most researchers are unprepared for project management, since it tends to be underemphasized in training programs. An epidemiologic project is not unlike running a business project with one crucial difference, the project has a fixed life span. This difference will affect many aspects of its management. Some areas of management that are affected are hiring, firing, evaluation, organization, productivity, morale, communication, ethics, budget, and project termination. Although the production of a study proposal raises many management challenges, if the proposal is approved and funds allocated, the accomplishments of the project are dependent more upon its management than any other factor.

A particular problem for investigators and staff, if they lack specific training or experience, is to fail to appreciate and prepare for the implications and exigencies of mass production.

1.2 The Data Management System

The data management system is the set of procedures and people through which information is processed. It involves the collection, manipulation, storage, and retrieval of information. Perhaps its most visible tool is the computer; however, this is merely one of many. Other "tools" are the instruments and data collection forms, the data management protocol, quality control mechanisms, documentation, storage facilities for both paper and electronic media, and mechanisms of retrieval. The purpose of the data management system is to ensure: a) high quality data, i.e., to ensure that the variability in the data derives from the phenomena under study and not from the data collection process, and b) accurate, appropriate, and defensible analysis and interpretation of the data.

---

* The original version of this chapter was written by H. Michael Arrighi, Ph.D.

1.3 Specific Objectives of Data Management

The specific objectives of data management are:

1.3.1 Acquire data and prepare them for analysis

The data management system includes the overview of the flow of data from research subjects to data analysts. Before it can be analyzed, data must be collected, reviewed, coded, computerized, verified, checked, and converted to forms suited for the analyses to be conducted. The process must be adequately documented to provide the foundation for analyses and interpretation.

1.3.2 Maintain quality control and data security

Threats to data quality arise at every point where data are obtained and/or modified. The value of the research will be greatly affected by quality control, but achieving and maintaining quality requires activities that are often mundane and difficult to motivate. Quality control includes:

- Preventing and detecting errors in data through written procedures, training, verification procedures, and avoidance of undue complexity
- Avoiding or eliminating inconsistencies, errors, and missing data through review of data collection forms (ideally while access to the data source is still possible to enable uncertainties to be resolved) and datasets
- Assessing the quality of the data through notes kept by interviewers, coders, and data editors, through debriefing of subjects, and through reviews or repetition of data collection for subsamples
- Avoiding major misinterpretations and oversights through "getting a feel" for the data.

Security concerns include: (1) legal, (2) safety of the information, (3) protection from external sources, (4) protection from internal sources. While abuse is more salient, accidental problems are more common. Typical preventive measures are removal or isolation of information that identifies research subjects (to protect confidentiality), redundancy, and backups (to protect against human and machine malfunction). The loss of important data due to failure to have a secure backup copy could be construed as negligence. Unfortunately, there can be an inverse relationship between security and accessibility/usefulness of the data.

1.3.3 Support inquiries, review, reconstruction, and archiving

Inquiries and requests for instruments and/or data may arise at any time during the project and after its completion. The funding agency will require a final report. Other investigators or interested parties (e.g., corporations whose products are implicated as health threats) may request a copy of the data set to pursue their own analyses. Rarely, an investigation may be conducted because of the salience of the findings, the involvement of parties with a large stake in their

implications, or suspicions or charges concerning the study. For example, Herbert Needleman, a pioneering investigator into the effects of childhood lead exposure on cognitive function, had his data and results audited by a scientific committee (which included a UNC faculty member). Proctor and Gamble, Inc., brought suit against the CDC to require the provision of data from their case control studies of toxic shock and tampons.

Concern about scientific misconduct and fraud continues to increase, and investigators have the responsibility to maintain documentation to allay any such charges should they arise. Increasingly, journals require that data (and supporting documentation) be retained for several years following publication. On a more mundane level, innumerable questions will arise during the course of the data analysis, and the project's data management system needs to be able to provide accurate and timely answers.

An important principle in data management, at all levels and stages, is the full accounting for data. Thus when a data collection activity takes place, there should be a detailed record of the number of subjects (if known) in the universe from which subject recruitment takes place and a complete tabulation within a set of mutually exclusive categories (dispositions). Typical dispositions are -- ineligibles according to the reason for their ineligibility (e.g., out of age range, medical conditions), nonparticipants according to the reasons for their nonparticipation (e.g., no telephone number, disconnected telephone, out of town, refused), participants whose data are excluded (e.g., too many missing data items, interviewer skeptical of participant's truthfulness), etc.

An audit trail is an essential mechanism to identify changes to the data at every step. The audit trail should document what changes were made, who made them, and where, when, and how the changes were made. Audit trails are important for responding to or recovering from: (1) legal challenges, (2) procedural issues, (3) minor problems, and (4) disaster.

Note that the above objectives apply to both manual and computerized systems.

1.3.4 Special issues in blinded studies

The HIV epidemic has led to a major activity in conducting blinded serosurveys to determine the prevalence of HIV infection in different settings, subgroups, and geographical areas. In order to avoid bias from nonresponse, a particular concern in HIV studies given the low prevalence of the virus in most populations and the fear and stigma associated with HIV infection and risk factors, methods have been developed for conducting blinded (unlinked) studies. Such studies use leftover blood collected for other purposes (e.g., medical tests) and are analyzed in such a way that identification of the individuals involved in the study is impossible. Under certain circumstances, such studies do not require informed consent, so that they can be free from nonresponse bias.

Special care is needed to design a data management system that can prevent the possibility of linking data to individual subjects. For example, standard data collection procedures such as the use of sequential ID numbers, inclusion of exact dates on all forms, and recording of supplemental information to clarify atypical data items can compromise anonymity. Indeed, unlinked studies

engender a basic conflict between the need to prevent linkage and major data management objectives, such as monitoring and quality control, which require the ability to be able to trace back and verify information.

1.4 The Components of Data Management

1.4.1 Management

The general concepts of management are as applicable to data management as they are to project management. Management issues are critical components of the data management system. The data are merely the objects being manipulated by the data management system. Unless there is adequate attention to the process, the data will not be worthy of much attention.

The investigative team is ultimately responsible for the outcome of the project. Even in those large projects where a project manager and a data manager are present, the investigative team are the project's board of directors. Management skills are required to evaluate the managers and ensure that they are doing a reasonable job, beyond the issue of is the project on schedule? Even for a relatively small project, researchers may need to work diligently to adapt to the managerial role, since many of the qualities that make for a good researcher are quite the opposite of those of a good manager:

Researcher	Manager
Optimal Solutions	Pragmatic Solutions
Accurate Solutions	Workable Solutions
Works with things	Works with people
Process Oriented	Outcome Oriented
Individually Successful	Group Successful

A good researcher requires creativity and may be considered a tinkerer, i.e., a person who is constantly changing things based on new ideas. A good manager is also creative but is less of a tinkerer. Constant change in a management situation results in confusion and a lack of consistency, which ultimately result in data of low quality. A few of the key management issues that directly affect the data management system are:

1.4.1.1 Two-way communication

Any person on the project can make a valuable contribution, comment, or observation. These contributions should be respected. Listening is an important aspect to learn what is truly happening and discover remedies. The person actually performing the task will often know more about the nuances and what is effective than anyone else. People have different degrees of readiness to express their ideas and concerns, and opportunities for informal one-on-one discussions (e.g., over coffee) are important.

1.4.1.2 Consistency

Consistency is essential in the implementation of the protocol, in the data collection process, and with regards to decisions made during the project. Lack of consistency may result from different decisions among the principal investigators, a lack of communication in relaying problems and solutions, and different decisions made on the same issue at different times. Innumerable minor questions (and many major ones) will arise during the course of a project. A useful perspective to use in resolving these questions is how would outside investigators view the decision. Decisions often need to be made quickly and recorded in some form that they can be readily consulted when the same or a similar question arises at a later date.

An inability to implement the study protocol consistently may result in selection bias or information bias. Information bias in confounding variables may compromise the ability to correct fully for confounding in the analysis. Also, when the methods and results are presented to others (a fundamental part of academic research) and the inevitable questions arise, it is very embarassing to have to describe and explain inconsistent methods.

1.4.1.3 Lines of authority and responsibility

Authority and responsibility need to be clearly defined and the designated persons accessible to the staff. Accessibility is often a problem in an academic research project where many of the staff are part-time employees, and the full-time management staff has other commitments resulting in a lack of accessibility to the staff. Among other things it is generally desirable to designate a specific person to be responsible for authorizing all changes to the computerized data sets.

1.4.1.4 Flexibility

The data management system must be flexible to respond to changes in the protocol, survey instruments, and staff changes. The longer a project runs the more susceptible it is to changes. Even a project using secondary data is subject to changes. Every project will undergo some modifications. Thus the data management system must be flexible to allow for easy modification.

1.4.1.5 Simplicity

Keep the data management system as simple as possible and within the talents of the (potential) staff. Simplicity reduces errors by reducing dependency on "key" personnel and by making the system easier to learn and implement.

Computers are wonderful tools in data management, but it is easy to complicate things by their use. The use of non-user friendly software packages or uncommon packages breeds complexity. Computerized systems actually increase the cost and technical support of systems. Their benefits are in the realm of increased efficiency and (hopefully) a reduction of errors. A small project may benefit from a predominantly, manual system when properly designed and implemented.

1.4.2 Integration

Integrate the data management system throughout the entire study process from the idea and proposal stage to the final paper, storage of information, and storage of data until planned destruction. Obviously, some concern to data management is given at the proposal stage during the budget and staffing process. More than this is needed; a general flow of the system should be thought through. This will provide a preliminary assessment of its resource demands and feasibility.

1.4.3 Standardization

Standardization extends not just to the instruments but to the procedures for participant enrollment, procedures for records review, data entry mechanisms, documentation, and other facets. This is essential to obtain quality information.

1.4.4 Pilot testing

Pilot testing is routinely done with the survey instruments. Rarely, does one pilot test crucial parts of the data management system. Use the pilot test of the survey instruments as an opportunity to pilot test aspects of the data management system, e.g. coordination of interviewers, call backs, coordination with other sources (participant identification) etc. A key aspect is to make this as real as possible to avoid the "test run" syndrome and a lack of seriousness among the staff.

The data management system may be pilot tested when the preliminary versions of the survey instruments are under evaluation and during the evaluation of the laboratory methods. If the project is sufficiently large, then a pilot test of the entire system may be done on the first few participants (5, 10, or 20). The project is then briefly halted for review and modification prior to complete implementation. Large projects make use of a "vanguard" cohort that goes through all aspects of the study sufficiently in advance of the actual study population to enable adjustment of instruments and procedures.

1.4.5 Quality Control and Quality Assurance

1.4.5.1 Redundancy

Duplication of data collection items is a well-established quality control procedure. This applies equally to survey instruments, laboratory procedures, and the data flow system. Duplication may occur in series or in parallel.

1.4.5.1.1 Parallel

Runs in parallel are the simultaneously evaluation of two data collection items. With a laboratory, this is the blind submission of two identical items for evaluation. With a survey instrument, this is the repetition of a question, perhaps in slightly altered format.

1.4.5.1.2 Series

Runs in series are the repetition of items at two different time points. With a laboratory, this is the blind submission of two identical items at different times. With a survey instrument, this is the repetition of all or part of the survey instrument at a different time. This may involve a call back of a selected subsample with a brief verification questionnaire asking some similar items. With data entry procedures, the verification of data entry is accomplished by the duplicate entry of the entire batch of items. These two entries are compared and non-matches are identified and re-entered. Double keying (also called key verification) though standard is not automatic, so must generally be specifically requested and budgeted.

1.4.5.2 Error introduction

A useful technique is to introduce errors into the data management system in order to evaluate the error trapping mechanism and consistency of error handling. This may be done by introducing erroneous data or identifying a particular problem and following it through the data management system.

1.4.6 Reduce the number of entry points

Each participant should enter the study the same way or each subject should have the same opportunity for entry. Different protocols should not be used to enroll different subjects of the same category (e.g. cases or controls, or the exposed or unexposed). This can be challenging when there are multiple sites.

Take a planned approach to every source of data, including logs, tracking forms, appointment systems. Hindsight can reveal usefulness of data sources not originally intended for analysis. Considerations in planning include design of the procedure for collecting, recording, coding, computerizing, verifying, ensuring security, and documenting. Try to limit situations where changes are made directly into a data base, with no audit trail. Without an audit trail it may be impossible to reconstruct data if an erroneous change is made or even to verify if a change has been made.

1.4.7 Monitor

Monitor the data management to ensure its proper implementation. For example, when a data collection activity is underway there should be frequent and regular review of the number of subjects participating, reasons for nonparticipation, problems encountered, etc. Data collection forms (or a sample, if the volume is large) should be scrutinized promptly to identify problems (for example, an excess of missing items, misunderstood items), for which corrective action may be possible. A manual or computerized system is important to keep track of data forms.

Crucial elements to identify are:

1. Adherence to the study protocol (to ensure the objectives of the protocol are maintained);

2. Consistency of the protocol's implementation;
3. Deficiencies (and strengths) of the data management system;
4. Response to changes, problems, crises - how well is the data management system detecting and responding to changes, problems, and crises? This monitoring may be accomplished through the use of erroneous data or by tracking the dates and items that were identified as problems, the date of their identification and the date of the correction.

1.4.8 Document

Documentation is a special challenge because of its lack of glamour and the fact that at any particular point in the study (before the end) competing urgent priorities make documentation a very hard activity to keep abreast of. Yet it is absolutely essential and cannot always be satisfactorily reconstructed after the fact. Budget time and personnel for documenting events, decisions, changes, problems, solutions. Review documentation as it is being developed to show the importance you assign to it and to make sure it is in the form you need it.

Document investigator meetings and committee meetings with each item on the agenda (1 or 2 sentences should suffice), followed by the decision or action (open, closed, resolved and summary thereof). Recounting all the discussion is nice but tends to be too lengthy - strive for succinctness and key points. Having minutes published within a day or two of the meeting forces them to be succinct and gets them done before memory fades. Keeping a contemporaneous journal or word processing document with (dated) notes about things that should be included in progress reports is another technique.

Project documentation should include:

- A succinct narrative of the objectives and methods of the study.
- A detailed chronology of events and activities through the life of the project, showing starting and ending dates and numbers of subjects for each data collection activity.
- For each data collection activity, a record of the detailed procedures used and an accounting of the number of subjects in scope, the number selected for data collection, and the disposition for all subjects (by category, e.g., could not contact, refusal). This material should have cross references to original sources (e.g., computer runs) to enable verification when necessary.
- Compendium of all data collection instruments, including documentation of the sources for questionnaire items obtained from pre-existing instruments. Results of pretesting and validation analyses or cross-references to them should be included.
- Lists and descriptions of all interventions applied and materials used (e.g., intervention materials, training materials)
- Documentation on all computerized data and final analyses (information on datasets, variables, computer runs - see below).

Obviously it will be easier to assemble these materials if careful documentation is prepared along the way. As a minimum every document should bear a date and a notation of authorship. Documents retained in a word processing file should if possible contain a notation of where the document file resides, so that it can be located for later revision or adaptation in creating related documents.

1.4.9 Poke around

The amount of activity and detail in a large project can easily exceed what the (usually limited) staff (and time-urgent investigators) are able to handle comfortably. Despite the highest motivation and experience, communication will be incomplete and important items will be overlooked. Investigators should review data forms regularly to familiarize themselves with the data in its raw form and verify that data collection and coding are being carried out as stipulated. It may also be worthwhile even to poke around occasionally in file drawers, stacks of forms, and among computer files.

1.4.10 Repeat data analyses

The fact that debugging is a major component of commercial software development suggests that investigators need to make provisions to detect and correct programming errors or at least to minimize their impact. One strategy is to have a different programmer replicate analyses prior to publication. Typically only a small portion of analyses that have been carried out end up in a publication, so this strategy is more economical than many others, though of course if serious errors misled the direction of the analysis much work will have been lost. Replication that begins as close as possible to raw data offers the greatest protection, but more often other methods are used to ensure the correctness of the creation of the first analysis dataset. An object lesson about the importance of verifying the accuracy of data analyses is offered by the following excerpt from a letter to the *New England Journal of Medicine* (Jan 14, 1999, p148):

> "In the February 5 issue, we reported the results of a study … We regretfully report that we have discovered an error in computer programming and that our previous results are incorrect. … After the error was corrected, a new analysis showed no significant increase …"

It is a good bet that error reports such as these are the tip of the iceberg.

Specification or programming errors can lead to irretrievable losses as well. In Cycle V of the National Survey of Family Growth (NSFG), for example, information about whether a pregnancy was wanted at the time of conception was not obtained for about five percent of pregnancies, due to errors in the computer-assisted personal interview (*Public Use Data File Documentation, National Survey of Family Growth, Cycle 5: 1995, User's Guide*, U.S. Department of Health and Human Services, PHS, CDC, National Center for Health Statistics, Hyattsville, Maryland, February, 1997). The first error occurred because a deliberate skip (respondents who had never had voluntary sexual intercourse [variable EVRHDVOL=2] were supposed to be skipped past the wantedness question) was operationalized based on EVRHDVOL≠1. That meant that women who had EVRHDVOL=blank were also skipped, which was a common situation since the EVRHDVOL question was not asked of women who had already said that their first intercourse was voluntary. The second error resulted from a combination of factors involving an intended skip past the

wantedness questions when a conception resulted from involuntary intercourse, the implementation of this skip based on a negative interval between dates of first voluntary intercourse and first conception (so that it appeared that conception occurred prior to first voluntary intercourse), and an estimation for the date of conception computed from the *month* when the pregnancy ended and the length of the pregnancy, which produced a result that could be a month before or after the actual date. Programming errors for other skip patterns caused 1,000 respondents to be mistakenly skipped around the miscarriage questions and 2,458 nonworking respondents to be mistakenly skipped around a series of questions about their most recent job. Since the NSFG interview included thousands of questions, this error rate is actually extremely low. Of course, if the affected items were the ones you were needing to analyze, you would still be disappointed – but probably not as much as the scientists and administrators whose careers were disrupted by the loss of the first Mars Lander due to a mix-up between English and metric units.

2. Data conversion

Picture stacks of questionnaires, stacks of medical record abstract forms, lists of laboratory results, photocopies of examination results, and the like. Before they can be analyzed, these original data need to be coded for computerization (even if the volume is small enough and the intended analyses are simple enough for manual tabulation, coding is still required). This process can be a major and arduous undertaking, and may involve the following steps for each data "stream":

1. Preparation of a coding manual stating the codes to be used for each data item and the decisions to be made in every possible situation that occurs (see sample coding manual);
2. Coding of a sample of data forms to pilot test the coding manual (see sample coded questionnaire);
3. Revision of the coding manual, re-coding of the sample, and coding of the remainder of the forms (see sample guidelines);
4. Maintenance of a coding log (see sample) recounting the identification number and circumstances for any data item about which a question arose or a nonroutine decision was made, to enable backtracking and recoding if subsequently indicated;
5. Recoding by supervisory personnel of a percentage (e.g., 10%) of data forms as a quality check.

Depending on the source of the data, coding can be largely routine (e.g., verifying that a response category has been circled and perhaps writing the appropriate code to be keyed) or highly demanding (e.g., evaluating a transcript of a medical record to determine the diagnosis and presenting complaint).

Coding is a good opportunity to review each data form for any irregularity that can be easily detected, including verbal information written on the questionnaire. Correction of irregularities (multiple responses to an item, inconsistent values, missing response, etc.) generally requires access to the data form, so it is easier to take care of the problem at the coding stage rather than when the forms have been filed and the computerized dataset is in use.

2.1 Data cleaning / editing – objectives

After coding, forms are data entered with some type of verification to detect and correct keying errors (double keying, manual proofing, etc.). The computerized data are then "cleaned" and "edited". Data cleaning and editing may also be referred to as "damage control." It is at this stage where the initial screening of the collected information is made to assess its validity and useful. Ideally, data cleaning is an ongoing process; initiated when the first results arrive. Detecting errors early may assist in minimizing them in the future and assist in their correction.

2.1.1 Incomplete data

Incomplete data are missing values for a single data item, incomplete or incorrectly completed instruments. An incorrectly completed survey instrument may be one that had the "skip" patterns improperly followed. The identification of these issues and their correction, when possible, are both of importance.

2.1.2 Extreme values

Extreme values for a variable are generally referred to as "outliers". Outliers may also occur for a site (in a multi-site study) or an interviewer who is more "extreme", with regards to timeliness, interview time, or responses. Outliers may meet one or both of two possible criteria.

2.1.2.1 Statistical

There are formal statistical tests for outliers. This process is designed to identify those values that may unduly influence a statistical analysis. A visual inspection of the data is actually quite informative in gaining impressions about potential outliers.

2.1.2.2 Substantive

For an outlier to be truly an outlier it must not make substantive sense, for example a hemoglobin of 0.5 (though a hemoglobin of 0.5 may not be a statistical outlier in a small group of patients with severe anemia, whose expected range might be between 3.5 and 8) or a height of 9 feet 10 inches with a given weight of 95 pounds in an apparently healthy adult.

2.1.3 Expected Results

Examining the data with an idea of what is expected is helpful in determining how "good" these data are.

2.2 Placement of the data editing process

Some reduction of the time and distance between data collection and entry into the analysis system is helpful in error correction. The editing of data should occur during all aspects of the data collection and analysis. Some editing may occur during or shortly after data collection; this often

involves manual means. Additional editing procedures will occur later during the formal coding and entry. Post-entry editing procedures will encompass the final aspect of the editing process.

2.2.1 Time of data collection

Examples of this are verification of respondents' identity, use of subject identification numbers with a check digit, clear marking of specimens with duplicate labels (including the caps), prompt reviewing of completed instruments, and provision of pre-labeled instruments.

2.2.2 Time of data entry (keying)

Many data entry programs enable checks for valid ranges and/or values as data are being keyed and can even include "hard" logic (consistency) checks. Modern large scale telephone surveys use computers to track and enter the data during the survey process. This ensures that the survey instrument is properly followed, responses are within tolerance range, and may even provide an opportunity for checks in consistency of response.

2.2.3 Post-entry

Most of the formal steps associated with data editing occur after the data have been keyed and verified. These involve the examination of individual records and their aggregates.

2.3 The steps in the Editing process

2.3.1 Manual Editing

As discussed above, manual checks are performed during coding of data forms. This stage checks for proper completion (skip patterns, etc.) of the questionnaire. Error correction may entail a return to the source of the original information. Or with an abstraction of medical (or other) records a photocopy of the source record may be obtained for comparison purposes.

2.3.2 Frequency distributions

SAS PROC FREQ (typically with the MISSPRINT or MISSING option selected) and PROC UNIVARIATE with the FREQ and PLOT options are useful in examining frequency distributions. Frequency distributions are helpful in identifying the extent and types of missing data, unusual patterns, and potential outliers. For example, birth weight in grams is generally recorded to the nearest ten grams, so the final digit should be zero. Blood pressure is generally recorded in mmHg, so the final digit should be uniformly distributed between 0 and 9. "The case of the missing eights" (Stellman SD, *Am J Epidemiol* 129:857-860, 1989) presents a case study of how an alert analyst noticed that a distribution of counts contained no counts with 8 as the least significant digit. Only after a great deal of checking and investigation was the problem traced to a programming error (a mixup of zero and the letter "oh").

2.3.3 Logic checks

These checks are evaluations of internal comparisons within a single observation.

"Hard" comparisons are made when there are obvious inconsistencies, for example pregnancy in males, or following the proper skip pattern of the survey instrument.

An example of a "hard" comparison is where sex is abstracted from two different sources. Those records with disagreement may be identified and handled according to the study protocol. This protocol may set those in disagreement to missing, or there may be preferred data source (e.g. respondent questionnaire versus medical record), or the original respondent may be contacted for verification. Disagreement across forms may reveal that the forms belong to different respondents.

"Soft" comparisons are possible but the exact values (cutpoints) will be study dependent. Marital status may be questioned if age at marriage is below a specified value, which may be gender-specific. The exact age is chosen dependent upon the population under investigation. Another check might include birth weight by gestational age.

2.3.4 Univariate displays

These statistics are useful in determining measures of central tendency (the familiar mean, median, and mode), measures of dispersion (standard deviation, variance, range, percentiles, skewness, and kurtosis). In addition, graphical representations (e.g. histograms, box plots, and normal probability plots) are helpful for describing data.

2.3.5 Bivariate displays

If the same data have been collected at multiple points, then agreement across the measurement points should be assessed. In addition, expected relationships can be examined: birth weight and gestational age, systolic blood pressure and diastolic blood pressure, height and weight. Unusual combinations should prompt examination for coding or entry errors.

Differences may be examined in the case of continuous variables and an acceptance protocol developed.

2.4 Treatment of missing values

Coding of missing and/or inconsistent responses merits careful thought. An item may have no response for a variety of reasons, and it is often useful to distinguish among those reasons. For example, an item may not be relevant for some types of respondents (e.g., a question about a prostate exam asked of a female participant, a question about the year in which injection drugs were last used asked of a participant who has not injected drugs), a "screening" question may have skipped the respondent around the item, the respondent may not recall the answer, the respondent may decline to answer, or the respondent may simply omit an answer (in a self-administered questionnaire) without giving a reason. An item such as "Have you had a Pap test in the past year? "

included in a self-administered questionnaire may not have been answered for any one of these reasons.

If there are many missing responses and only a single missing value code is used, then a frequency distribution will often leave the analyst wondering about the usefulness of the item, since if there are too few analyzable responses the item conveys limited information. It is preferable to use a different missing value code for each situation. Then a frequency distribution can show the number of responses of each type.

2.4.1 Techniques for coding missing values

A widely-used convention for coding missing values uses out-of-range numeric codes, such as "999", "998", or "888", to represent missing values, with the number of digits equal to the field length. For example, in the public use datafile for the National Survey of Family Growth, responses of "don't know" are coded 9, 99, 999, or 9999; refusals are coded 8, 98, 998, or 9998; and "not ascertained" values are 7, 97, 997, or 9997, depending on the column length of the original data items. There are several limitations to this coding procedure. First, statistical packages may not recognize these as missing values, so that it is easy (and embarrassing) to have the actual values actually included in analyses. Second, it may be necessary to use different codes for the same missing reason due to different field lengths. Third, the numbers provide no useful mnemonic device.

One very useful facility in the Statistical Analysis System (SAS) is the provision for special missing value codes. A missing value for a numeric variable can be coded with one of 27 different missing value codes, consisting of a dot or a dot followed by a single letter. Although it is rare to use more than two or three for a given variable, an analysis could differentiate among "not applicable", "does not know", and "refused" by coding these as .N, .K, and .R, respectively. A more elaborate coding for a variable such as menstrual discomfort might differentiate among not applicable due to gender (.M), not applicable due to having had a hysterectomy (.H), not applicable due to natural cessation of menses (.C).

These values can be tabulated separately in frequency distributions, so that the extent and nature of "missingness" for a variable can be quickly assessed, and the analyst can keep track of denominators more easily. SAS generally does not include missing values coded in this way in calculations, which saves some programming effort and protects against programming lapses. The TABLES statement in PROC FREQ provides an option (MISSING) to treat missing values the same as all other values (useful to examine percentages of missing values) and an option (MISSPRINT) to display missing values in tables but not include them in the denominator for percentages (permitting the correct computation of percentage distributions for analysis while permitting easy verification of the number of data points and reasons for their absence.

2.5 Outliers

Examination for extreme values (range checks) is also a crucial preliminary step in the screening of data. First, outliers should be checked to the original data forms to verify accuracy of

transcription. If the outlier cannot be dismissed as an error, then care must be taken to avoid distorting the analysis.

2.5.1 What to do with them?

Outliers may be replaced with a missing value, but then the observation is lost with regards to the analysis (and in a mathematical modeling procedure, the entire observation is unused). Moreover, if the outlier is a legitimate value, then simply deleting it is a questionnable procedure.

The analysis can be repeated with and without the outlier data to assess the impact of the outlier on the analysis. Or, the analysis can be repeated using (nonparametric) statistical procedures that are not affected by outliers and the results compared to parametric procedures - or nonparametric procedures can be used completely. Such procedures typically involve medians, rather than means, or focus on the ranks of the values of a variable rather than on the values themselves. Categorical procedures in which a variable is first categorized into groups will often be unaffected by extreme values.

2.6 Concern with Data Cleaning / Editing

There is a problem with the handling of these missing values, outliers, and other edit checks: more attention is given to the extreme problems and less attention to those errors that are not as visible. For example, a transcription error resulting in a birthweight of 2,000 grams being recorded as 3,000 grams may go completely undetected once data entry is completed. But this misclassification bias may have a substantial effect if it occurs in a large enough proportion of observations. The entire data management system should be designed to minimize and reduce these errors. Related to this concern is the comparison with "expected" values. While this is a useful tool in inspecting and understanding the data, there is a concern in trying to force the data into an expected distribution. Thus the focus is on those errors of the extreme. An error in the opposite direction, from more extreme to less, is missed by this definition of data examination. This latter concern applies equally through the remainder of the data examination and analysis.

2.7 Documentation

Documentation covers all aspects of the data as well as all problems identified, their solutions, and all changes made to the data. Some techniques are:

- Keep a master copy of the questionnaire and record changes and decisions for each item. Cross index this questionnaire with the variable names in the computer files.
- Keep at least the original data and all programs leading to creation of the final dataset, so that any intermediate dataset may be recreated if need be. (This is the rationale for not making direct changes to the data base.)
- Document computer programs with a unique identifier (i.e., program name), title of project, brief description of the program, input and output, any dependencies (programs

that MUST be run prior to this or essential data bases), date of request and person, date of development and analyst, including modifications).

- Document computer programs within the program (in comment statements and title statements), files and programs, and externally (i.e. notebooks).
- Maintain a notebook of program runs in chronological order, showing the (unique) program name, date run, programmer, history (e.g., rerun of an earlier version), data set used, and one-line description. Sometimes programs that create datasets are listed in a separate section from programs that analyze datasets.
- Try to use self-documenting methods. Adopt a system of naming conventions for datasets, computer runs, and variable names. Choose meaningful variable names if possible, and allow for suffixes (e.g., using only 7 characters in a SAS variable name leaves the 8th character for designating recodes of the original variable). Assign internally-stored labels to data sets and variables (SAS LABEL or ATTRIB statement). If more than 40 characters are needed, add a comment in the program that creates the variable or dataset. Consider using value labels (formats) to document the values for each variable.

3. Data analysis

With the availability of microcomputer statistical packages, it is easy to compute many statistics that previously required the assistance of someone with biostatistical training (and with fewer distractions from the task of data analysis), with an increase in the danger of uniformed, inappropriate, or incorrect use of statistical tests (W. Paul McKinney, Mark J. Young, Arthur Hartz, Martha Bi-Fong Lee, "The inexact use of Fisher's Exact Test in six major medical journals" *JAMA* 1989; 261:3430-3433).

The first stages of data analysis should emphasize obtaining a "feel" for the data, i.e., some familiarity with their essential features. The process of examining the data to understand them is integrated throughout the cleaning and analysis. Always question data and examine them with a critical view. The same concepts used in data cleaning and editing are applicable in trying to understand the data. Specifically, these are the expected values, missing values, and outliers. Now they are applied in a "multivariate sense."

Many of the methods of approaching a dataset are similar to those described above under data cleaning, such as examination of:

1. Univariate distributions (frequency distributions [PROC FREQ], summary statistics [PROC UNIVARIATE], graphs [PROC UNIVARIATE, PROC PLOT or other].
2. Crosstabulations (frequency distributions across important groupings, such as sex, race, exposure, disease, using PROC FREQ)
3. Scatterplots showing pairs of continuous variables
4. Correlation matrices

These analyses should include the assessment of agreement where it is expected to occur. It is often helpful to prepare summary tables of basic information from the above examination, that can be used for reference purposes during later stages of analysis and writing.

3.1 Data reduction:

Data reduction is an essential activity that, like data management, takes place at virtually every place where data are involved. In the data analysis phase, data reduction involves deciding whether and how continuous variables can be grouped into a limited number of categories and whether and how to combine individual variables into scales and indexes. There is also the need to derive conceptually more meaningful variables from individual data items.

3.2 Graphical representation

There are many graphical packages available that provide the ability to plot, view, and to an extent analyze data. Graphical representations of data are extremely useful throughout the examination of the data. Statisticians are often familiar with these techniques for examining the data, describing data, and evaluating statistical tests (e.g. plots of residuals). The visual impact of a graph is informative and will increase the understanding of the data and limit the surprises that may occur. There are few general principles, as each data set is different and will have an individual approach. Many of the modern statistical graphics packages available on personal computers have a variety of functions such as fitting curves, for example, linear, quadratic, other polynomial curves, and spline curves.

3.3 Expected values

Perhaps, the single most important concept to remember is to have an idea of what is expected. This concept has been applied during the editing and cleaning process. Understanding what is expected is a function of both the study design and the values of the parameters in the target population. For example, if randomized allocation has been used, then the randomized groups should be similar. If controls are selected from the general population via random digit dialing methods, then their demographics should reflect the population as a whole. When examining a table, first check the variables, labels, and N's for the total table and the subcategories that are not included to make sure that you understand the subset of observations represented. Second examine the marginal distributions to make sure they conform to what you expect. Then examine the internal distribution, particularly, with regards to the referent group. Finally proceed to assess the association or other information in the table.

3.4 Missing values

The impact of missing data is magnified for analyses involving large number of variables, since many analytic procedures require omitting any observation that lacks a value for even one of the variables in the analysis. Thus, if there are four variables, each with missing data for 10% of the observations, in a worst-case situation 40% of the observations could be omitted from the analysis.

To assess the extent and nature of missing data for a variable, a complete "missing value" analysis should ideally be done. That means comparing the presence/absence of information for a variable with other key factors, e.g. age, race, gender, exposure status, and/or disease status. The goal is to identify correlates of missing information. Relationships are indicative, though not conclusive, of selection bias. This analysis may give insights into how to impute values for those missing (e.g., missing cholesterol could be estimated as a function of sex, age, race, and body mass). Strong relationships between one covariate and missing values for another indicate that imputed values should be stratified by levels of the first covariate.

Although they receive relatively little attention in introductory treatments of data analysis, missing values are the bane of the analyst. Examination of the data for missing values (e.g., via SAS PROC FREQ or PROC UNIVARIATE) is an essential first step prior to any formal analyses. Special missing value codes (see above) facilitate this examination. Missing values are a serious nuisance or impediment in data analysis and interpretation. One of the best motivations to designing data collection systems that minimize missing values is experience in trying to deal with them during analysis!

3.4.1 Effects of missing data

Two kinds of missing data can be distinguished: data-missing and case-missing. In the former case, information is available from a study participant, but some responses are missing. In case-missing, the prospective participant has declined to enroll or has dropped out. This discussion will address the situation of data-missing.

Missing data have a variety of effects. As a minimum, missing data decrease the effective sample size, so that estimates are less precise (have wider confidence intervals) and statistical tests have less power to exclude the statistical null hypothesis for observed associations. This problem is compounded in multivariable analyses (e.g., stratified analysis or logistic regression), since most such procedures drop every observation which has a missing value for any of the variables in the analysis. Thus, a logistic model with eight variables can easily lose 30% of the observations even if none of the individual variables has more than 10% missing values.

In both univariate and multivariable analyses, missing data leads to what might be referred to as the problem of the "changing denominator". Each one-way or two-way table may have different numbers of participants, which is both disconcerting to readers and tedious to keep explaining. One workaround is to analyze only complete-data cases (i.e., observations with no missing values), but the price in number of observations lost may be unacceptable.

Missing data situations are characterized in terms of the degree and patterns of "missingness". If there is no systematic pattern to missing data for a particular item, i.e., all participants are equally likely to omit a response, then the missing values are missing completely at random (MCAR). When data are MCAR, then estimates from the nonmissing data will not be biased by the missing data, since the nonmissing data is essentially a simple random sample of the total (potential) data.

It is probably more often the case that different groups of participants have different rates of missing data. In this case, the data are missing at random (MAR) (assuming that missing data occur randomly within each group). If groups who differ in their rates of missing data also differ in their distributions of the characteristic being measured, then overall estimates of that characteristic will be biased.

For example, if persons with multiple sexual partners are more likely to decline to answer a question on that topic, then the estimate of the mean number of partners or the proportion of respondents with more than X partners will be biased downwards. Estimates of associations with other variables may also be distorted. Furthermore, attempts to control for the variable as a potential confounder may introduce bias (from selectively removing observations from the analysis) or due to incomplete control for confounding.

3.4.2 What to do about missing data?

As in so many other areas of public health, prevention is best. First, data collection forms and procedures should be designed and pretested to minimize missing data. Second, it may be possible to elicit a response from a hestitant or unsure respondent (but such elicitation must avoid the hazards of eliciting an inaccurate response or contravening in any way the participant's right to decline to answer), to recontact participants if questionnaire review turns up missing responses, or to obtain the data from some other source (e.g., missing information in a hospital medical record may be available from the patient's physician). Third, it may be possible to combine data from different sources to create a combined variable with fewer missing values (e.g., sex from a questionnaire and sex from an administrative record, though the issue of differential accuracy of the sources may be an issue).

Despite the best efforts, however, missing data are a fact of life, and it is the rare observational study that avoids them completely. Nevertheless, the smaller the percentage of missing data, the smaller a problem they will create and the less it will matter how they are dealt with during analysis.

3.4.3 Do not try to control for missing values of a confounder

The suggestion arose some years ago to treat missing values as a valid level of a variable being controlled as a potential confounder. For example, if an association was being stratified by smoking, there might be three strata: smoker, nonsmoker, smoking status not known. Recent work suggests that this practice may actually increase confounding and is not recommended.

3.4.4 Imputation for missing data

In recent years a great deal of work has gone into developing analytic methods for handling missing data to minimize their detrimental effects. These methods seek to impute values for the missing item responses in ways that attempt to increase statistical efficiency (by avoiding the loss of observations which have one or a few missing values) and to reduce bias that results when missing data are MAR, rather than MCAR (i.e., missing data rates vary by subgroup).

One simple method of imputation, now out of favor, is simply to replace missing values with the mean or median of the available responses. This practice enables observations with missing values to be used in multivariable analyses, while preserving the overall mean or median of the variable (as computed from the nonmissing responses). For categorical variables, however, the mean may fall between categories (e.g., the mean for a 0-1 variable may be .3), and for all variables substituting a single value for a large number of missing responses will change the shape of the distribution of responses (increasing its height at that value and reducing its variance), with effects on statistical tests. Moreover, if missing values are not MCAR, then the mean of the observed values may be biased and therefore so will the mean of the variable after imputation.

3.4.5 Randomized assignment of missing cases

A more sophisticated approach is to draw imputed values from a distribution, rather than to use a single value. Thus, observations without missing values (complete data cases) can be used to generate a frequency distribution for the variable. This frequency distribution can then be used as the basis for randomly generating a value for each observation lacking a response. For example, if education was measured in three categories -- "less than high school" (25% of complete data cases), "completed high school" (40%), or "more than high school" (35%) -- then for each observation with education missing, a random number between 0 and 1 could be drawn from a uniform distribution and the missing value replaced with "less than high school" if the random number was less than or equal to 0.25, "completed high school" if the number was greater than 0.25 but less than or equal to 0.65, or "more than high school" if greater than 0.65.

This method avoids introducing an additional response category and preserves the shape of the distribution. But if the missing data are not MCAR, the distribution will still be biased (e.g., greater nonresponse by heavy drinkers will still lower the estimate of alcohol consumption; greater nonresponse by men may also lower the estimate of alcohol consumption).

3.4.6 Conditional imputation

Modern imputation methods achieve more accurate imputations by taking advantage of relationships among variables. If, for example, female respondents are more likely to have a confidant than are male respondents, then imputing a value for "presence of a confidant" can be based on the respondent's sex. WIth this approach, confidant status among men will be imputed based on the proportion of men with a confidant; confidant status among women will be imputed based on the proportion of women with a confidant. In this way, the dataset that includes the imputed values will give a less biased estimate of the population values than will the complete-data cases alone.

A simple extension from imputation conditional on a single variable is imputation conditional on a set of strata formed from a number of variables simultaneously. If the number of strata is too large, a regression procedure can be used to "predict" the value of the variable to be imputed as a function of variables for which data are available. The coefficients in the regression model are estimated from complete-data cases.

Imputed values are then randomly assigned (using a procedure such as that outlined above) using the stratum-specific distributions or predicted values from the regression model. This strategy provides superior imputations for missing values and preserves associations between the variable being imputed and the other variables in the model or stratification. The stronger the associations among the variables, the more nearly accurate the imputation. There does remain, though, the problem of what to do when the value of more than one variable is missing. If in actuality two variables are associated with each other, then imputing values to one independently of the value of the other will weaken the observed association.

3.4.7 Joint imputation

Yet another step forward is joint imputation for all of the missing values in each observation. Picture an array which categorizes all complete-data observations according to their values of the variables being considered together and a second array categorizing all remaining observations according to their configuration of missing values. Suppose there are three dichotomous (0-1) variables, A, B, C and that A is known for all respondents but B and/or C can be missing. The arrays might look like this:

Complete data cases

						% distribution conditioning on			
Stratum #	A	B	C	Count	Percent of total	A	A & B	A, C=0	A, C=1
1	0	0	0	400	33	53	67	80	
2	0	0	1	200	17	27	33		75
							100		
3	0	1	0	100	8	13	67	20	
4	0	1	1	50	4	7	33		25
						100	100	100	100
5	1	0	0	240	20	53	62	83	
6	1	0	1	150	13	33	38		88
							100		
7	1	1	0	40	3	9	67	17	
8	1	1	1	20	2	4	33		12
						100	100	100	100
Total				1,200	100				

Missing value configurations

Configuration	A	B	C	Count
a.	0	0	.	12
b.	0	1	.	18
c.	1	0	.	10
d.	1	1	.	30
e.	0	.	0	40
f.	0	.	1	10
g.	1	.	0	15
h.	1	.	1	25
i.	0	.	.	20
j.	1	.	.	10

In this example, the eight strata in the cross-classification of the complete data cases are numbered 1 through 8, and the percentages for each stratum are computed in four different ways: unconditionally (i.e., the count as a percentage of all of the complete-data cases), conditionally based on the value of A only, conditionally based on the value of A and B, and conditionally on the value of A and C [the latter requires two columns for clarity]. Meanwhile, the 10 possible missing data configurations are arrayed in the second table and labeled a. through j.

Imputation is then carried out as follows. Missing value configuration a. has A=0 and B=0, so the 12 cases in this configuration belong in stratum 1 or stratum 2. To preserve the distribution of the complete data cases in those two strata (67% in stratum 1, 33% in stratum 2 – see column headed "A & B"), the 12 cases are randomly assigned to stratum 1 and stratum 2 with assignment probabilities in that proportion, so that stratum 1 is expected to receive 8 and stratum 2 to receive 4. The 18 cases in configuration b. have A=0 and B=1, so they belong in either stratum 3 or stratum 4. These 18 cases will be randomly allocated between these two strata with probabilities proportional to the distribution of the complete data cases across those two strata (which happens to be the same as the strata with A=0 and B=0). Configurations c. and d. will be handled in the same manner. Configuration e. has A=0 and C=0, so the 40 cases in this configuration belong in either stratum 1 or 3. These 40 cases will be randomly assigned to strata 1 or 3 in proportion to the distribution in the column headed "A, C=0". Thus the random assignment procedure will on average assign 32 cases (80%) to stratum 1, and 8 cases (20%) to stratum 3. The remaining configurations will be handled in the same manner. Configuration i. has A=0 but no restriction on B or C, so the 20 cases in this configuration will be randomly allocated across strata 1, 2, 3, or 4 according to the distribution in the column headed "A" conditional on A=0.

Joint, conditional, imputation makes maximum use of the available data on the three variables, adjusts the distribution of each variable to give a better estimate of that expected for the population as a whole and preserves many of the two-way associations involving variables being imputed. The procedure can be carried out using a modeling procedure instead of a cross-classification, which enables the inclusion of more variables.

Model-fitting using the EM ("Expectation Maximization") algorithm is the current state of the art. The BMDP AM procedure uses this algorithm, but it is designed for continuous variables

with a multivariate normal distribution and imputes each variable independently, so that two-way associations are weakened. A new program by Joe Shafer at Pennsylvania State University uses the EM algorithm with categorical variables and jointly imputes data; however, it requires very powerful computer resources.

3.4.8 Multiple imputation

All of the above procedures result in a single dataset with imputed values in place of missing values. However, since the imputed values are derived from the rest of the dataset, analyses based on them will understate the variability in the data. As a corrective, the imputation process can be carried out repeatedly, yielding multiple datasets each with a (randomly) different set of imputed values. The availability of multiple imputations enables estimation of the additional variance introduced by the imputation procedure, which can then be used to correct variance estimates for the dataset as a whole.

[With thanks to Drs. Michael Berbaum, University of Alabama at Tuscaloosa and Ralph Foster, Research Triangle Institute (NC USA) for educating me on this topic and reviewing this section.]

3.5 Outliers

Outliers are now examined with respect to a multivariate approach, i.e., are there any extreme values? For example, you stratify the exposure - disease relationship by a factor with 4 levels. The observation is made of the 4 stratum specific odds ratios of 2.3, 3.2, 2.7, and 0.98. The fourth stratum indicates a potentially strong interaction. What if this stratum contains only 6 observations? Even though the association may be statistically significant, collapsing the strata is reasonable as the most extreme table may be a result of imprecision. Alternatively, the values of the most extreme table may be recategorized.

3.6 Creation of analysis variables

The variables defined to contain the data in the form it was collected (as responses on a questionnaire, codes on a medical abstraction form, etc.) do not always serve the purposes of the analysis. For example, a questionnaire on risk behaviors might use separate items to ask about use of crack, injected cocaine, injected heroin, and snorted heroin, but a single variable combining these behaviors ("yes" if used cocaine or heroin, "no" if used neither) might be more useful for the analyst. In that case a derived variable would be created (treatment of missing values becomes an issue here, as well). Similarly, a question about marital status and a question about living with a "romantic partner" might be combined into a variable indicating "living with spouse or partner".

3.7 Deciding which values to include in analyses

It is not always clear which values to include in analyses. For example, generally missing values are excluded from the denominator for computation of percentages, except when the purpose is an assessment of the extent of missing data. Sometimes, however, it is more meaningful

to treat at least some categories of missing values in the same way as non-missing values. For example, a series of items about specific changes in behavior might be preceded with a screening question, such as "Have you make changes during the past year to reduce your risk of acquiring HIV?"

If the respondent answers "yes", s/he is asked about specific changes; otherwise the specific items are skipped. In this situation, a missing value due to the skip really means "no". This situation can be handled by creating a new variable for each of the specific items or by recoding the existing variables to "no" when the screening item was answered with "no", or by other techniques. In contrast, a "true missing" would be present if the individual item was not answered even though the screening question was answered "yes". This "true missing" would probably be excluded from the analysis. Similarly, if this type of behavior change was not relevant for the respondent, then the item is "not applicable" and the observation would probably be excluded as well ("probably", because the appropriate treatment depends upon the purpose of the analysis and intended interpretation).

3.8 Assessment of assumptions

During this stage, the assumptions underlying the statistical techniques are assessed. For example, a chi-square test has certain minimum expected cell sizes. A t-test assumes a Gaussian (normal) distribution in the population. Other assumptions are those made about reality. For example, what if a person responds to the question on race by circling 3 responses, Black, Hispanic, and White. There is a study protocol to classify such an individual; however, this protocol may differ from other similar studies or the U.S. Census, or state birth certificates, etc. This may have an impact on the expected distribution and /or interpretation.

3.9 Examination of study questions

Data analyses may be approached in an exploratory fashion or in pursuit of answers to specific questions. Ideally the latter should have been specified in the research proposal or well before the analysis process has begun. Often new questions (or all questions) are formulated during the analysis process. In either case, it is highly desirable to articulate specific questions as a guide to how to proceed in the data analysis.

Besides their relevance to the questions at hand, analyses generally need to reflect the study design. For example, cross-sectional designs do not provide direct estimates of incidence, matched designs may warrant matched analyses.

Bibliography

Calvert, William S. and J. Meimei Ma. *Concepts and case studies in data management* . Cary, NC: SAS Institute, c1996.

Davidson, Fred. Principles of statistical data handling. Thousand Oaks, California, SAGE, 1996, 266pp.

Graham JW, Hofer SM, Piccinin AM. Analysis with missing data in drug prevention research. IN: Collins LM, Seitz LA (eds). Advances in data analysis for prevention intervention research. NIDA Research Monograph 142. U.S. D.H.H.S., N.I.H., National Institute on Drug Abuse, 1994, 13-63.

Marinez, YN, McMahan CA, Barnwell GM, and Wigodsky HS. Ensuring data quality in medical research through an integrated data management system. Statistics in Medicine 1984; 3:101-111.

Hybels, C. Data management outline. Presented at the American Geriatrics Society Summer Workshop. 1989.

Hse J. Missing values revisited. Presented at the all-Merck statisticians conference, October 23, 1989.

Hulley, Stephen B. and Steven R. Cummings. Designing clinical research: an epidemiologic approach. Baltimore, Williams & Wilkins, 1988. Chapter 15: Planning for data management and analysis.

Meinert, Curtis L.; Susan Tonascia. Clinical trials: design, conduct, and analysis. New York, Oxford, 1986.

Raymond, Mark R. Missing data in evaluation research. Evaluation & the health professions 1986;9:395-420.

Spilker, Bert; John Schoenfelder. Data collection forms in clinical trials. Raven Press, 1991.

Websites

Reasearch Data Management, Joachim Heberlein, University of Minnesota, August 28, 1999 has links and case studies concerning 1. federal, University, and disciplinary guidelines governing ownership, access, and retention of research data; 2. choices, decisions, and justifications regarding data (a) accuracy and reliability, (b) ownership, (c) access, (d) use, and (e) retention; 3. preparation of guidelines for the management of research data. www.research.umn.edu/ethics/modResearch2.html

Appendix

```
*************************************************************************;
*  The following SAS code can be adapted for use to create a check character
*  for ID numbers which can then be used to detect transcription errors when
*  the ID numbers are read again.  For example, numeric ID numbers can be
*  generated by any system and then suffixed or prefixed with a check character.
*  The ID's can be printed on questionnaire labels, specimen labels, coding
*  forms, or other data collection or tracking instruments.  When the ID numbers
*  are keyed along with the associated data, the data entry program can use
*  an adaptation of the code that follows to verify the accuracy of
*  transcription of the ID number itself.
*
*  The check character generated by the following code will detect transcription
*  errors involving a misrecording of any single digit of the ID number,
*  reversal of any two digits, or even multiple errors, with relatively rare
*  exceptions.  Since errors in ID numbers can be among the most troublesome
*  to detect and correct, use of a check character is recommended.
*
*  The code on which this SAS routine was based was developed by
*  Robert Thorton at the Research Triangle Institute (RTI), based in turn
*  on an article by Joseph A. Gallian ("Assigning Driver's License Numbers",
*  Mathematics Magazine, February 1991, 64(1):13-22).  Thornton's code forms
*  the method for creating and checking ID numbers for the RSVPP study.
*  This SAS version was developed by Vic Schoenbach, 10/18/94, 10/24/94;
*
*  Here are some sample ID's and their corresponding check digits,
*  taken from a list of ID numbers provided by RTI to Project RAPP:
*
*         5-1120 -> S  (i.e., the complete ID number is 5-1120-S)
*         5-1111 -> T
*         5-1101 -> W
*         5-1011 -> A
*         5-1001 -> D
*         5-1002 -> B
*         5-2001 -> V
*         5-3001 -> Q
*
*************************************************************************;
*;

* This program reads a list of ID numbers and assigns check characters.
  The program also reads the check characters assigned by RTI so that these
  can be displayed alongside the calculated check characters to facilitate
  verification of the correctness of the calculated check characters,
  for testing purposes;

data;   * Create a SAS dataset with the original and calculated numbers;

  * Do not write the following variables into the dataset:      ;
    drop alphabet char1 lng sum i mod23 complem ;

* Define three variables:  the ID number, the check digit, & a 1 byte work area;
attrib strng length=$22 label='ID number needing check digit';
```

```
attrib ckd length=$1 label='Check digit to be calculated';
length char1 $1;  * For picking out one character at a time from the ID;
length sum 8;  * For calculation purposes;
alphabet= 'ABCDEFGHIJKLMNOPQRSTUVW';

infile cards;  * Input data file will be on "cards" (i.e., right after
                                                        the program);
input strng $  rti_ckd $ ;  * Read in data (consisting of ID's and the
                  RTI check digit, so that it can be printed in the output);

sum=0;  * Temporary variable to compute running sum;
lng=length(strng);  * Get length of ID to be processed;
if lng > 21 then do;  * Check that the ID is not too long;
        file print;
        put // '*** Error: ' strng= ' is too long = (' lng ')' //;
        file log;  return;  end;

do i = 1 to lng;    * Iterate through each digit of ID number ;
  char1 = substr(strng,lng-i+1,1);   * Extract a character from the ID;
                             * (Hyphens will be ignored);
  if char1 ^= '-' then
     if char1 < '0' or char1 > '9' then do;  * Must be a valid digit - if not
                                               then print error message;
       file print;
       put // '*** Error: Non-numeric character in ID: ' strng= char1=  //;
       file log; return;   * Go back for next ID number;
       end;   * End of then do;
     else do;  * (To get here, character must be a digit from 0-9);
       sum = sum + ((i+1) * char1);  * Take the sum of the digits of the ID
                                  number, weighting each digit by its position;
       end;   * End of else do;
end;    * End of do i = 1 to lng;

* Weighted sum has been obtained - now reduce it;
mod23 = mod(sum,23);  * Calculate the remainder after dividing by 23;
complem = 23 - mod23;  * Take the complement from 23;
ckd=substr(alphabet,complem,1);  * The check character is the
                                   corresponding letter of the alphabet;
return;

cards;   * Here come the test ID's -- note that one is invalid;
   5-1120  S
   5-1111  T
   5-11R1  W     (invalid ID number)
   5-1101  W
   5-1011  A
   5-1001  D
   5-1002  B
   5-2001  V
   5-3001  Q
run;               *(end of list of ID numbers);

* Display the results to verify correctness;
  proc print;  var _all_;  run;
```