Quality Control and Good Epidemiological Practice

I.13

Preetha Rajaraman, Jonathan M. Samet

13.1 Introduction

The use of data is fundamental in epidemiology. Epidemiologic research on causation uses data in a search for the true nature of the relationship between exposure and disease. Similarly, research on the consequences of interventions seeks an unbiased characterization of the effects of independently varying factors on the outcome measure(s). One of the most rewarding moments for a researcher is obtaining the preliminary results from his or her study. However, the question "do I believe what I see?" should immediately come to mind. The answer to this question is determined in large part by the more mundane but critical question of how good is the quality of the data, rather than by the elegance of the scientific method. Errors that occur during study population selection or in the measurement of study exposures, outcomes, or covariates can lead to a biased estimate of the effect of exposure on risk for the disease of interest. Misclassification of exposure or disease that occurs randomly between all study participants decreases the power of the study to detect an association where it exists. Data collection that is differentially biased may have more severe consequences, and can lead to an incorrect assessment of the relationship between exposure and disease.

The inherently important issue of study quality is becoming of even greater consequence as the findings of epidemiological studies gain in impact, and the field of epidemiology gains wider acceptance as an essential element of biomedical research (Samet 2000; Samet and Lee 2001). Results of epidemiological studies are routinely reported in the media, receiving widespread attention because the findings have evident relevance to the populace. Epidemiologic evidence is also used to inform regulatory and legislative policy making (Goldman 2001). The decision to set airborne standards for particulate matter in the United States, for example, was largely fueled by evidence from epidemiological studies (Greenbaum et al. 2001). Epidemiology often figures prominently in litigation, where the study methodology can become a point of debate (Bryant and Reinert 2001; Goldman 2001). Given the significance of epidemiologic evidence for decision-making, the results of epidemiological studies often face close scrutiny and questions may be raised about every aspect ranging across data quality, study methods, study conduct, data analysis and interpretation of findings.

Even if external questioning and auditing are not anticipated, the researcher nonetheless faces the responsibility of assuring the quality of the study and preventing the widespread dissemination of misleading or incorrect information. For example, findings from several cohort studies on air pollution and mortality figured prominently in a 1997 decision by the U.S. Environmental Protection Agency (EPA) to promulgate a new standard for airborne particulate matter. The great weight given to the data by the EPA led to a call for access so that others could check and analyze the data. An extensive re-analysis of the data was carried out, including validation of elements of the original data as well as replication and extension of the original analyses by an independent group (Krewski et al. 2000; Samet et al. 1997). The controversy surrounding the use of data from the air pol-

lution cohort studies eventually led to a Congressionally-mandated requirement for sharing data with policy implications that have been collected with federal funds.

Many hypotheses of current interest in epidemiological studies call for the incorporation of data from multiple centers and involve collection of data from large populations according to centrally standardized protocols. Data sharing has also become more common, and approaches to doing so for larger grants in the United States have been mandated by the National Institutes of Health. In order to enhance statistical power, data from individual studies are often pooled, or summary results are combined using meta-analysis. These approaches to data utilization place a further demand for meticulous study documentation so that data from a study are readily usable by persons other than the original investigators.

General methods have long been available for assuring the quality of data. The idea of creating a high quality end-product using process improvement initially emerged in the context of industrial business models. Early efforts at delivering quality products to customers were based on inspecting products at the end of a factory line and eliminating those products that did not meet standards ("quality control"). The idea of improving all procedures that affect the quality of the manufactured products ("quality assurance") represented a fundamental shift in paradigm for industrial manufacture. Incorporating quality considerations into the process rather than the product has since gained widespread acceptance in the business and engineering communities (International Organization for Standardization 2003).

Although there is a vast literature on quality control in general, the issue has not received much formal attention in the epidemiological setting. Within epidemiology, much of the writing on data quality and good epidemiological practice is focused on the conduct of clinical trials (Canner et al. 1983; Cooper 1986; Dischinger and DuChene 1986; DuChene et al. 1986; Gassman et al. 1995; Hilner et al. 1992; Knatterud et al. 1998; Meinert and Tonascia 1986; Neaton et al. 1990; Vantongelen et al. 1989). While clinical and laboratory guidelines can easily be modified to make them more applicable to observational studies, few sources specifically address quality issues for the most common epidemiological study designs. In an early attempt to bridge this gap, the Epidemiology Task Group of the Chemical Manufacturers Association (CMA) compiled a set of guidelines for good epidemiology practice for occupational and environmental epidemiologic research (Cook 1991; The Chemical Manufacturers Association's Epidemiology Task Group 1991). An overview of data quality issues for epidemiological studies is also provided by Szklo and Nieto (2000) and by Whitney and colleagues (Whitney et al. 1998). Methods to improve data quality in medical registries are reviewed by Arts et al. (2002).

This chapter provides a general overview of data quality and guidelines for good practice in epidemiological research. The fundamental premise is that quality considerations should be integrated into every phase of the study from initial hypothesis formulation to the final publication of findings and archiving of data.

Obtaining data completely free from error clearly would be prohibitively expensive, and often impossible. The goal is therefore not error-free data, but rather planning and implementing cost-effective procedures that guarantee the validity of the primary results to an acceptable degree. The epidemiological researcher needs to be able to gauge the extent of any errors, and assess the consequences for interpretation of data analyses. The idea of "quality control" versus "quality assurance" is carried over from the industrial management literature into the epidemiological literature, with a distinction made between activities that take place prior to data collection (quality assurance), and activities that occur during and after data collection to correct data errors (quality control).

The ubiquitous nature of quality issues, both in terms of where these issues can arise and how they affect study results can be captured by an extended metaphor. In an article describing the causation of bias, Maclure and Schneeweiss present the idea of an "Episcope" through which an epidemiologist views a putative association between a causal agent and morbidity. Just as a user of a large telescope would be skeptical about whether and how image degradation exists, an epidemiologist should think about how and why an observed association between exposure and disease might be biased (Maclure and Schneeweiss 2001). A similar idea can be applied to data quality. As published study results are viewed through a "Datascope," a discerning epidemiologist should be wary of how the final image (the published results) may have been distorted by quality considerations during the design, conduct and dissemination of the study. Working backwards, the observer might ask a string of questions, such as "Were the observed results more likely to be published because they were positive findings? Based on the analysis, were published inferences appropriate? Were the methods of analysis suitable? Were data keyed in correctly? Has the data been collected appropriately? Was an appropriate population defined?" Each of these questions points to one or more study quality issues. Using the metaphor of the datascope, we will highlight the main issues regarding study design and conduct, and present ways in which to improve epidemiologic practice and data quality.

13.2 The Datascope

Imagine, for a moment, that published study results can be viewed only through a large telescope. As you peer into the lens, the initial picture is barely discernible. On your right is a panel with focusing controls. The first dial allows you to adjust out any distortion caused by publication bias. When you optimize this dial, the image becomes slightly clearer. The next control allows you to tune out faulty inferences. Again, you turn this knob to make the image somewhat clearer. The process continues until the results are finally sharply focused. Although we do not literally look through a telescope every time we view the results of a study, we are in fact looking at an association that may well be "out of focus" depending on how well the study was designed, conducted and interpreted. Errors in the

measurement of the exposure, outcome, or other covariates can be thought of as unfocused datascope controls that contribute to degradation of the final image.

Let us consider, in some more detail, the datascope controls that manipulate sources of measurement error. The farthest dials from the observer are located in the planning phase of a study and influence purely "quality assurance" activities. For a more in-depth discussion of the planning stage see Chap. I.12 of this handbook. Errors occurring at the study planning stage are summarized below.

- *Errors in study conception*
  If the study rationale and design are not carefully formulated, the rest of the study could be rendered completely irrelevant. Errors in study conception include inadequate literature review, consideration of an inappropriate study design, and failure to plan the validation of exposure or outcome variables.
- *Errors in the selection, design, or procedures for use of instruments measuring exposure*
  The instrument selected for study exposure measurement might not cover all sources of the active agent. Conversely, the measurement instrument might include sources of exposures that are not biologically relevant, or measure exposure for a time period that is etiologically unimportant. In survey instruments, the phrasing of questions or instructions could lead to misunderstanding or bias (cf. Chap. I.10 of this handbook). Insufficient detail in the protocol for instrument use or inadequate consideration of a standardized method for dealing with unusual situations can lead to collection of poor quality data.
- *Inadequate training of study personnel*
  Even if study procedures are very well defined, inadequate training of data collection staff in the application of these procedures can introduce errors in the data (cf. Chap. I.10 of this handbook).

The next set of controls is activated during the conduct of a study and includes activities that generally fall under the categories of quality assurance as well as quality control. For instance, validation studies of instruments and equipment ensure that collected data will be accurate (quality assurance), but can also be used to correct errors in data (quality control). Sources of exposure measurement error that can occur during data collection are described in detail elsewhere (Armstrong and White 1992) and summarized below.

- *Improper execution of the study protocol*
  Errors related to study protocol execution include the misinterpretation of, or deviation from, standard operating procedures by study technicians. Mistakes in interpretation of the study protocol often arise from poor clarity of the manual of operations or inadequate training of study personnel. For example, if the standard operating procedure states that a fasting blood glucose level should be measured but does not specify the time required to have elapsed after the last meal, the interpretation of "fasting" may differ from technician to technician. Errors in data can also result from improper handling of biologic specimens, or the failure of subjects to read or understand instructions in self-administered questionnaires.

- *Errors related to study participants and intra-individual variability*
  Subjects may have poor recall of past exposures, or allow recent exposure to influence their memory of past exposure. Individuals also tend to over-report socially desirable behaviors such as exercise, and underreport socially undesirable habits such as smoking. Additionally, short-term variability in the biological characteristics of a subject can lead to unrepresentative measurement of exposure or outcome. For example, differences in the level of an exposure biomarker measured at a specified time after exposure are likely to be due partially to individual differences in metabolizing the agent of interest.
- *Changes in the accuracy of measurements over time*
  Failing to standardize and recalibrate laboratory equipment is likely to introduce data drift as calendar time progresses. In long-term studies, the instrument used for measurement may change over time, and the agent of interest in biological specimens may be subject to degradation. Also, as the study personnel get more experienced through the course of the study, changes in the handling of procedures and instruments may occur.
- *Mistakes in data processing*
  Data that are recorded inaccurately, illegibly, or incompletely are very difficult to correct after the fact. Transcription of the data to electronic files introduces more chances for error, both within a study site, and between field sites and the data coordinating center. At the coordinating center, programming or procedural errors may corrupt the database or modify data inappropriately. Errors can also be introduced by undocumented changes or modifications to a local or central database.

The final panel of controls on the datascope, closest to the observer, consists of purely "quality control" dials, which influence study quality after the data have been collected. Examples of these errors are presented below.

- *Inappropriate data analysis*
  If data analysis is not preceded by familiarization with the nature of the data, the chosen analyses may not be appropriate. Specifying the wrong model for analysis, for instance, can lead to completely erroneous results and inference.
- *Poor reporting of data*
  Omitting the results of important data analyses, or presenting unnecessary information can obfuscate the study results. Lengthy, verbose explanations and poorly labeled graphs and figures add to the confusion. Inappropriate inference given the study results can also be misleading.

In order to achieve the highest quality data possible, each of the sources of error described in the planning, design, conduct, and conclusion of a study should be minimized. Conceptually, this can be thought of as turning the appropriate datascope dial to obtain the best image possible.

A review of the different sources of error that can occur during study planning, design and conduct informs the datascope user as to where he or she can affect final data quality. The ultimate goal is to optimize the datascope dials in order to

minimize error and achieve the clearest possible picture of the study results. In the rest of this chapter, we present aspects of quality control and good epidemiological practice that can reduce data error. The chapter will follow the same organization as the datascope control panels, beginning with the planning phase of a study, moving onto quality considerations during study conduct, and finally describing activities that occur after data collection. Where applicable, the working of the datascope will be illustrated using the example of measuring blood pressure in a hypothetical study whose main research question is whether elevated blood pressure leads to increased risk of coronary disease.

Quality Considerations in the Planning Phase

13.3

Protocol

13.3.1

The development of a comprehensive study protocol is essential to good epidemiological practice. The *study protocol* is a narrative document that describes the general design and procedures used in the study. It can be distinguished from the study *manual of operations* (Sect. 13.3.3) by its generality and absence of specific details for day-to-day study conduct. The study protocol assists the staff in understanding the context in which their specific activities occur. A well-designed study protocol can, and should, guide all aspects of the study. In general, a protocol would include the following sections: a short descriptive title; a description of performance sites and personnel; a description of background and significance; results of preliminary studies; study design and methods; a time line for completion of major tasks; ethical considerations, and references. Quality assurance and quality control should be addressed in each relevant section of the protocol, and also summarized in a separate section. Although restrictions or recommendations provided in the guidelines for research grants applications for the U.S. National Institutes of Health may not be applicable to grants funded through other mechanisms, these guidelines nevertheless provide useful suggestions for creating study protocols (U.S. Department of Health and Human Services 2001). Recommendations for protocol write-up are also included in the Guidelines for Good Epidemiology Practices for Occupational and Environmental Epidemiologic Research (The Chemical Manufacturers Association's Epidemiology Task Force 1991). The typical sections of a study protocol are summarized in Table 13.1 (see also Chap. I.12 of this handbook).

Improving the Datascope Image by Choosing Appropriate Measures of Hypertension

In the planning phase of the study, investigators should make provision for collection appropriate measures of hypertension. While clinicians favor

the diagnosis and treatment of hypertension in terms of diastolic blood pressure elevation, data from the Framingham Study in Massachusetts indicate that systolic blood pressure is a better predictor of disease outcome (Kannel 2000). Additionally, ambulatory blood pressure can be measured with an automated device so that multiple measurements can be made across the course of typical activities. Studies show that such recordings provide information predictive of disease risk beyond that obtained with measurements made at a single assessment (Clement et al. 2003).

Table 13.1. Guidelines for preparation of a study protocol*

Section	Guidelines for good epidemiological practice
Title	Descriptive and to the point.
Names, titles, degrees, addresses and affiliations of the study director, principal investigator, and all co-investigators	Possible conflict of interest should be identified and resolved.
Name(s) and address(es) of the sponsor(s)	Possible conflict of interest should be identified and resolved.
Proposal abstract	Informative and succinct.
Proposed study tasks and milestones	Timetable should be realistic and identify possible sources of delay.
Statement of research objectives, rationale, and specific aims	Clearly state the purpose of the investigation, describe whether the study will be hypothesis-generating or hypothesis-testing, and whether the study will confirm previous findings or result in new findings.
Critical review of the relevant literature	Include animal, clinical, and epidemiological studies. Do not restrict search to electronic databases (e.g. PUBMED, TOXLINE), older articles might be missed. Describe the occurrence of exposure and outcome variables. Identify potential confounders and effect modifiers. Identify gaps in current knowledge.
Description of the research methods	Describe the overall research design, and why it was chosen. Consider alternative designs.
	Define exposure and outcome variables, and identify data sources for these and other variables of interest. Check whether the measure of exposure represents the biologically active agent and etiologically important time period.
	Calculate the projected study size and statistical power (if appropriate).
	Describe procedures for collecting data.

table to be continued

Table 13.1. (continued)

Section	Guidelines for good epidemiological practice
	Provide a detailed description of the methods of analysis.
	Define how exposure and outcome variables will be categorized for analysis.
	State how confounders and effect modifiers will be treated in the analysis.
	Outline the major strengths and limitations of the study design.
	Provide criteria for interpreting the study results, including ways of assessing statistical, clinical, and biological significance.
Description of plans for protecting human subjects	Describe risks and benefits of participating in the study.
	If appropriate, provide plans for obtaining informed consent.
	Describe procedures for maintaining confidentiality of subjects and data.
Description of quality assurance and control	Describe for all phases of the study.
Resources required to conduct the study	Detail the expected time, personnel, and equipment required for the study.
Bibliographic references	Include all relevant references.
Addenda, as appropriate	Examples of useful addenda include copies of collaborative agreements, institutional approvals, informed consent forms, and questionnaires.
Dated protocol review and approval sign-off sheet	Document dated amendments to the protocol.

* adapted from the Guidelines for Good Epidemiology Practices, Epidemiology Task Group (The Chemical Manufacturers Association's Epidemiology Task Force 1991).

Documentation of Operations and Procedures 13.3.2

The consistency and validity of study data are greatly enhanced by the establishment and application of *standard operating procedures* for routine data collection tasks (a *standard operating procedure* is defined here as a standardized method or process for conducting a routine research procedure). If standard procedures have been well described, variability is likely to be much lower across study sites, interviewers, or technicians. Uncorrected variability introduced by interviewers or technicians can decrease study power.

Standard procedures should be clearly described for all study procedures, including (but not limited to) raw data collection, coding of death certificates, assessment of error rates, and management of archived data. Each standard operating

procedure should state the purpose of the procedure, provide a detailed description of the procedure including forms and equipment to be used, and either designate the person responsible for the procedure, or explain what training will be needed (The Chemical Manufacturers Association's Epidemiology Task Force 1991). Detailed quality control and quality assurance guidelines for the collection of laboratory samples are provided in the U.S. Toxic Substances Control Act (TSCA) standard for Good Laboratory Practices (US Environmental Protection Agency (EPA) 1989).

Once the various standard operating procedures are established, they should be integrated and summarized in the form of a study *manual of operations*. The *manual of operations* is a document or collection of documents that completely describes the procedures used in a study center. Developing a *study handbook*, which contains a series of tables, charts, figures, and specification pages that outline the main design and operating features of a study (largely without the use of a written narrative) is a useful first step in the development of the manual of operations, and can also act as a quick reference for study personnel. The study protocol, handbook, and manual of operations should be reviewed for clarity and completeness.

Since the initial version of the manual of operations is almost certain to contain some errors, pre-testing of the manual prior to finalization is essential. All aspects of the study protocol should be tested on a population similar to the one that will be studied, including the administration of surveys, sending of samples to laboratories, and the generation of and response to quality control reports. Refinements to the protocol that are identified from the pilot study can be incorporated into the final study manual of operations.

Improving the Datascope Image Using Standard Operating Procedures

Inter- and intra-technician variation in blood pressure readings viewed under the datascope can be reduced by clear and detailed descriptions of the method of measurement. Application of a standard operating procedure can also reduce variability in blood pressure measurement within a subject. Specifying details such as how the study participant should be seated, which arm the cuff should be applied to, and how long the study participant should remain quiet before the reading is taken can reduce the influence on the study measurement of factors that affect an individual's blood pressure.

13.3.3 Personnel, Training and Certification

Integral to study conduct is the availability of personnel with the necessary education, experience and training to perform assigned functions. The planning stage of the study is the appropriate time to consider personnel requirements, what kind of training will be necessary, and how often training should occur. Job descriptions should be written for each individual who will be supervising or engaging in the

conduct of the study. For jobs that require training, procedures for initial and re-training of personnel should be established. Re-training may be necessary if substantial time has elapsed since the initial training, if a technician is found to be introducing a systematic error into the data, or if the study protocol changes. For each of the study personnel, a summary of relevant training and experience, including study certification and recertification, should be maintained and kept up-to-date.

Consistency in the training of personnel across sites improves comparability of data collection across different study sites. This training can be centralized, or site-specific. Often, a combination of both approaches is used (see Sect. 13.4.1 for more detail). Study personnel should be required to follow standard operating procedure. If training will be difficult or time-consuming, it is prudent to train at least two individuals for each task in case one of the trained technicians leaves the study. Certification standards should be set, and might include completion of a specified number of tests for key procedures, including some under observation.

Aside from the obvious benefit of consistency in data collection, training study personnel also increases the interviewers' or technicians' perceived value of the data that are being collected. This may influence the amount of care taken in following the protocol. Some studies use computer instruction, video cassettes, or teleconferencing to reduce the costs associated with training. While the use of computer or video training is convenient, these methods lack some of the benefits of face-to-face training, such as the opportunity for staff members to share ideas, and the opportunity for scientific presentations that remind personnel of the importance of their work (Whitney et al. 1998).

Improving the Datascope Image by Training and Certification

Some of the variation in blood pressure measurements viewed under the datascope could arise if a technician measured blood pressure in a different way each time he or she took a measurement, or if different technicians had different ways of reading the same measurement. One way to minimize these sources of error and improve the datascope image is to train and certify study technicians. In the MRFIT (Multiple Risk Factor Intervention Trial), technicians were trained in taking blood pressure measurements using training tapes and a double stethoscope (Dischinger and DuChene 1986). The training tapes consisted of two recordings of the Korotkoff sounds for twelve subjects. The first tape of Korotkoff sounds was used for training, and the second for testing. A video training film that presented twelve blood pressure readings, with sufficient time to determine and record systolic and diastolic blood pressure after each reading, was also used. Finally, supervisors and trainees took simultaneous measurements of three subjects using a double stethoscope. The differences in the readings of the trainer and technician had to fall below a certain criterion for the trainee to pass. Technicians were certified after completing the training tapes, passing a written test

on procedures for taking blood pressure measurement, and passing the double stethoscope test. Recertification was required at regular intervals, or if examination of collected data indicated that a technician had a bias with respect to other technicians in a clinic.

13.3.4 Data Collection Forms and Instruments

Exposure and outcome measures for epidemiologic studies can be collected in a variety of ways. Methods of data collection include mailed self-administered questionnaires, interviewer-administered questionnaires, measures of blood or other tissues, physical measures, medical tests, use of medical or exposure records, or sampling for environmental contaminants (White et al. 1998). Most studies use more than one method of data collection.

The use of data that have been collected already ("secondary data") has the key advantage that the data already exist. Studies using secondary data are thus likely to be more cost and time-efficient than studies with primary data collection. Sources of secondary data, such as population-based registries, often allow for a much larger sample size, and can be more representative of the general population (Hearst and Hulley 1988). A substantial disadvantage of using existing data, however, is that the collected data may not adequately address the particular research question of interest. An additional drawback is that the method of collection and the quality of the secondary data are not under the researcher's control. For this reason, researchers using secondary data should carefully review data documentation and evaluate the quality and validity of these data to the extent possible (Clive et al. 1995; Gissler et al. 1995; Goldberg et al. 1980; Horbar and Leahy 1995; Maudsley and Williams 1999; Sorensen et al. 1996; Wyatt 1995). For details on the use of secondary data see Chap. I.4 of this handbook.

Most epidemiological studies collect some or all of their data using phone, mail, or self/interviewer-administered questionnaires. Data from such questionnaires, however, can be subject to various sources of bias. For instance, study participants filling out a self-administered questionnaire might report socially acceptable rather than strictly accurate results. Moreover, ways of responding to the survey may differ between participants in the study, depending on factors such as the age, gender, or racial/ethnic group of the participant. Conversely, participants may respond differently to interviewers of different age, gender or ethnic background. For further details see Chap. I.10 of this handbook. Multi-center studies encounter the additional problem of differences in data collection between study centers. In long term studies, these biases can change over time. Smoking, for example, is generally less socially acceptable today than it was 20 years ago in the United States and consequently more likely to be underreported (Ling and Glantz 2002). The acceptability of smoking also varies by ethnic groups, which may in part explain the fact that African-American high school seniors are far less likely to smoke than are white seniors (Wallace, Jr. et al. 2002). Measurement error that occurs

because of the use of a survey instrument can be minimized by careful design and pre-testing of the survey, and the application of standardized interviewing techniques.

The main objective of survey design is to allow the efficient collection of data that are valid, reliable, and complete. Standardizing forms within a study is important for internal validity. Consistency of forms across studies allows more meaningful comparison with other studies, and also makes the study results more generalizable. Both internal and external form standardization can be achieved by the use of pre-existing validated study instruments. Examples of validated questionnaire instruments include the American Thoracic Society questionnaire to assess respiratory symptoms (Comstock et al. 1979), and the Willett food frequency questionnaire (Willett et al. 1985). If a validated instrument is not readily available, several sources in the literature provide guidelines for questionnaire design to maximize clarity and ease of administration. These include recommendations for physical format, as well as instructions on how to word the text of instructions and questions (Dillman 1978; Hosking et al. 1995; Knatterud et al. 1998; Meinert and Tonascia 1986; Wright and Haybittle 1979a, b, c). Studies that enroll participants of different ethnic groups may need to accommodate different languages by using interpreters, or by having translated versions of the questionnaire. However, a question might change subtly upon translation, and data generated from different languages may not be entirely comparable. For this reason, independent back-translation of questions to the original language is strongly recommended. An example of the need for back-translation is provided by data from a health survey which showed lower data reliability of data for Hispanics interviewed in Spanish than for Hispanics interviewed in English when no back-translation was done. An independent back-translation aimed at creating a linguistically equivalent version to the Spanish version indicated several instances in which the two versions were idiomatically different and appeared to have affected the seriousness with which the interview situation was perceived, in turn leading to response discrepancies (Berkanovic 1980).

Pre-testing of the survey instrument on a population similar to the study population allows the detection of flaws in the survey design and instrument before full-scale data collection begins. Separate analysis of pre-test data by language version, for example, might identify problems in translation. In the Hypertension Prevention Trial, which was designed to test the effectiveness of changes in dietary intake of calories, sodium, and potassium, a test cohort of 78 participants was enrolled, and used for the testing of forms and procedures. Data that were generated from the test cohort were used to identify problems in survey design and collection, and were not analyzed with results from the main study (Prud'homme et al. 1989).

Accuracy and consistency are also important for laboratory or clinical equipment. The study should be planned so that all study personnel and sites begin by using identical equipment. In anticipation of measurement drift over time, procedures to maintain and recalibrate equipment should be established. In the Sleep Heart Health Study (Quan et al. 1997), overnight sleep data were collected

from subjects using a portable monitor. Sites were notified to have the monitor evaluated and procedures assessed when less than 85% of results scored by the monitor were of "good" or better quality (Whitney et al. 1998). Standard and random zero sphygmomanometers for blood pressure measurement in the MRFIT study (Kjelsberg et al. 1997) were maintained and calibrated according to a regular schedule, and subject to standard checks at least every other month in the case of the standard sphygmomanometer, and every week in the case of the random zero instrument (Dischinger and DuChene 1986).

As more advanced technology becomes available to measure an exposure or outcome, there may be justification to update study equipment. In such cases, data should initially be collected using both the old and new equipment to establish the comparability of the two instruments, since a change may introduce subtle differences that are only apparent as substantial data are collected using the new approach.

13.3.5 Planning Response Rate

In order to curtail the possibility of bias and increase the generalizability of study results, it is important to achieve the highest response rate possible (Gordis 2000; Wacholder et al. 1992). A recent systematic review of 292 trials found that factors which more than doubled the odds of response to surveys were: the inclusion of a monetary incentive with the questionnaire, designing surveys to be of more interest to participants, and the use of registered mail (Edwards et al. 2002). Other factors which have been reported to increase response rate are shorter questionnaire length (Dillman 1978; Eaker et al. 1998; Hoffman et al. 1998; Kalantar and Talley 1999; Kellerman and Herold 2001; Little and Davis 1984; Martinson et al. 2000; Spry et al. 1989), personalizing questionnaires (Maheux et al. 1989), using colored ink (Edwards et al. 2002), contacting participants before sending questionnaires, providing stamped return envelopes (Choi et al. 1990), and using written or telephone reminders (Asch et al. 1997). Questionnaires originating from universities are more likely to be returned than questionnaires from other sources, whereas surveys eliciting information of a sensitive nature are less likely to be returned.

While the use of a monetary incentive is probably the factor that has been shown most consistently to increase response rates (Gibson et al. 1999; Gilbart and Kreiger 1998; Hoffman et al. 1998; Kellerman and Herold 2001; Martinson et al. 2000; Parkes et al. 2000; Perneger et al. 1993), increasing the amount of the incentive results in diminishing returns of questionnaires after a certain point (Halpern et al. 2002; James and Bolstein 1992; Spry et al. 1989). In the United States, the $2.00 bill seems to be a cost-effective monetaryincentive (Asch et al. 1998; Doody et al. 2003; Shaw et al. 2001). Making a pre-payment of the incentive appears to be more cost-effective than promising payment on completion of the questionnaire (Schweitzer and Asch 1995). Including the monetary incentive in the first mailing rather than in subsequent mailings has resulted in higher response rates (John and Savitz 1994). Non-monetary incentives, while reported to increase

response rates over having no incentive, do not appear as effective as monetary incentives (Kellerman and Herold 2001; Martinson et al. 2000).

Contact rates generally tend to be lower for individuals who are young, male, black, of lower socio-economic status, or employed full-time (Collins et al. 2000; Cottler et al. 1987; Moorman et al. 1999). In the context of a case-control study, response rates are often lower for controls (Moorman et al. 1999). Even within control groups, different types of controls have different response rates. For example, in the United States, controls chosen from Health Management Organizations have been shown to have a higher response rate than controls drawn from lists of licensed drivers (Slattery et al. 1995).

Long term cohort studies, in addition to having to address response rates to study questionnaires, also face the issue of loss to follow-up. The loss of cohort members to follow-up is conceptually similar to response rate, in that loss to follow-up can constitute an important source of selection bias and also limit external validity. Participants may be lost to follow up either because they drop out of the study of their own volition, or because the study investigators lose track of them. As with other types of epidemiological studies, loss to follow-up can lead to reduced study power and may result in biased estimates of risk. Strategies for minimizing loss to follow up include pre-enrollment screening of participants for willingness to participate in a long-term study, collecting names of personal contacts and proxies for participants, maintaining regular contact with study participants, using incentives for remaining participants, and maintaining tracking systems to follow participants (Hunt and White 1998). One must keep in mind, however, that populations comprised of volunteers are usually different from the population as a whole. In general, measures of relative risk are less affected by the lack of external generalizability than measures of absolute or attributable risk.

Validity and Reliability 13.3.6

The absence of bias in data measurement is called *validity*, or accuracy. The precision, or reproducibility, of collected data is known as *reliability*.

Validity Studies

The capacity of a measure to capture the true value of the exposure, outcome, confounder, or modifier of interest in the study population is known as its validity. While it is desirable to obtain the most accurate measurements possible of exposure or outcome, such measurements usually come at the price of increased cost, invasiveness, or time involvement. When faced with these constraints, epidemiologists often choose to collect less accurate measures of exposure.

The accuracy of the study's main method of exposure measurement can be assessed using validation studies which compare the study exposure measure to a more accurate measure of exposure ("gold standard"), either in a sub-sample of study participants or in a different population. For instance, evidence of validity

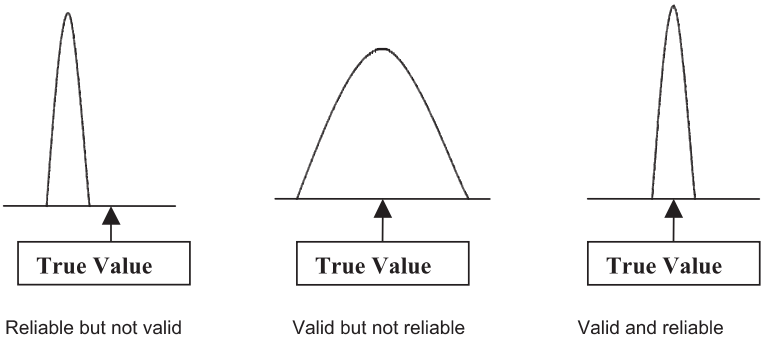

Figure 13.1. Graphs of hypothetical test results illustrating the distinction between validity and reliability

can be provided by comparing study estimates of an environmental exposure to industrial hygiene measurements or biomarkers of exposure (Cherrie and Schneider 1998; Cherrie et al. 1987; Dosemeci et al. 1997; Hawkins and Evans 1989; Kipen et al. 1989; Kromhout et al. 1987; Tielemans et al. 1999). The comparison of reported nutrient intake on a questionnaire with a biochemical indicator provides another example of this approach (Ascherio et al. 1992; Johnstone et al. 1981; Post and Kromhout 1991; Sacks et al. 1986; Willett et al. 1983).

The establishment of a serum pool can facilitate validation of biological sample processing in the study. Study measurements can be compared with results from a "gold standard" external laboratory. If study measurements deviate randomly from the gold standard, the study result would be attenuated towards the null hypothesis. However, if deviations from the gold standard are found to vary according to the presence and level of important variables such as follow-up time, or the exposure or outcome of interest, the study results may be biased.

Data from validation studies can, additionally, be used to account for uncertainty in the data analysis. Measurement error correction models can be developed that use validation study data to adjust the full data for measurement error (Holford and Stack 1995; Rosner et al. 1992; Spiegelman et al. 1997; Stram et al. 1999; Thompson 1990). In the Framingham Study (Dawber et al. 1951), for example, a small validation study was conducted to estimate the relationship between the surrogate measurement (food frequency) and the "true" measurement (diet record). Based on information from the main study relating the surrogate to disease outcome, and information from the validation study relating true and surrogate exposure, corrected point estimates of risk were calculated (Spiegelman et al. 1997). For a general discussion of statistical methods to account for measurement errors see Chap. II.5 of this handbook.

While validation studies may help form a clearer picture of the true relationship between exposure and outcome, such studies are not without their own limitations. For one, the gold standard used for comparison may itself be subject to error,

thus the term "alloyed" gold standard (Wacholder et al. 1993). While calibration methods for such alloyed gold standards have been described, these complex models cannot be applied in all situations (Kaaks et al. 2002; Spiegelman et al. 1997). A second limitation of validity studies is that participants in these studies are not always representative of all participants. Subjects who volunteer to take part in a validity study are likely to be more compliant than non-volunteers would have been. Additionally, feasibility constraints often limit validity studies to small sample sizes, which can lead to statistical imprecision.

Reliability Studies

Data variation can arise within study participants (*biological variability*), or due to variation in exposure assessment or physiological measurements introduced by study technicians. Blood pressure within an individual, for example, experiences short-term changes due to factors such as activity and mood. Different blood pressure measurements taken on the same individual are thus likely to vary for physiological reasons regardless of how accurately these measurements are made. Study technicians can add an extra component of variation to the measurements, either because a given technician reads a measurement in slightly different ways each time (*intra-observer variation*), or because different technicians read a measurement in different ways (*inter-observer variation*). Variability can also be introduced as samples degrade over time.

As illustrated in the paragraph above, variability in data can arise due to true change, measurement error, or random biological variation. The component of variability in which the researcher is most interested is the true change in study exposure that might influence the outcome under consideration. The separation of desired variability in the data (true change) from undesirable variability due to measurement error or random (biological) variation can be partially assessed by incorporating into the main study a series of sub-studies that are designed to assess the reliability of the study data (cf. Chap. I.11 of this handbook).

Reliability studies can be used to assess various components of variability, such as the comparability of measurements taken by: the same technician at a given visit; different technicians at a given visit; the same/different technician at different points of time, or the same/different technician using different instruments. Inter-observer and intra-observer variability can be assessed using a set of calibration samples that are read several times by each technician, and processed by multiple technicians. Biological variability can be assessed by having a single technician perform repeat studies on a subset of participants, although some amount of variability assessed in this way would be due to technician variability (Whitney et al. 1998). For durable data, such as X-ray films or dietary recall records, variation over time can be assessed by comparing evaluation of the same samples at different times in the study. In instances where samples are limited or perishable (e.g. blood or urine), a pre-selected set of "quality control" specimens should be set aside at the beginning of the study so that small amounts of these specimens can be periodically submitted for processing. Technicians handling quality control samples should be

unaware that the samples are different from other study samples being processed, in order to prevent differential handling.

Reliability studies which collect replicate measurements at the same point of time are useful in the identification of possible data errors, as well as in the calculation of more accurate measures of exposure. Averaging repeated measures has been recommended as an effective method of decreasing the measurement error associated with a single measurement (Armstrong and White 1992; Canner et al. 1991; Holford and Stack 1995; White et al. 1998).

When validity is reported, the number of samples that are deemed unacceptable for analysis should be stated, since this may indicate a bias in the remaining samples. The examination of whether reliability estimates differ according to relevant characteristics such as exposure, confounding factors, or outcome allows some assessment of whether differential misclassification is occurring in the data.

Improving the Datascope Image by Obtaining a More Valid Exposure Measure

Using the average of three blood pressure measurements taken on a study visit could result in a clearer picture of the individual's blood pressure than would a single blood pressure measurement.

Measures of Agreement

Quantifying the agreement between two different methods of measurement requires the use of some measure of agreement. The choice of statistic depends on the type of variables being compared, and the purpose of the comparison (Table 13.2). The calculation of different measures of agreement, and advantages and disadvantages of each measure have been reviewed by Szklo and Nieto (2000).

The basic measures of validity for binary categorical variables are sensitivity and specificity, for which the study value of the exposure or outcome is compared to the "true" value, measured by a more accurate method (Example 1). The *sensitivity* of a test is the ability to correctly identify those individuals who have the disease or exposure characteristic of interest. The test *specificity* is the ability to correctly identify those individuals who do not have the disease or exposure characteristic of interest. A limitation of the use of sensitivity and specificity is that very few diagnostic tests are inherently dichotomous. Most diagnostic tests are based on the characterization of individuals based on one or more underlying traits, such as blood pressure or serum glucose level. Values for the sensitivity and specificity would vary according to the cut-off level used to separate "diseased" (or exposed) from "undiseased" (or unexposed) individuals. In addition, if measurement error occurs, individuals with true levels of the underlying trait close to the test point are more likely to be misclassified. Since the distribution of underlying traits also determines disease prevalence, sensitivity and specificity can vary from population to population (Brenner and Gefeller 1997).

Example 1. *Calculation of Sensitivity and Specificity*

		Gold Standard Results		
		Positive	Negative	Total
Study Results	Positive	a	b	$a + b$
	Negative	c	d	$c + d$
		$a + c$	$b + d$	
	Sensitivity = $a/(a + c)$			
	Specificity = $d/(b + d)$			

Agreement for categorical variables (e.g., X-ray readings by radiologists) is generally reported using variations of the percent agreement and kappa statistics. While overall percent agreement is intuitive and easy to calculate (Example 2), it can make agreement look artificially high, since there is likely to be considerable agreement between two observers reading negative, or normal, results. An alternative approach is to disregard subjects labeled as negative by both readers, to calculate the percent positive agreement (Cicchetti and Feinstein 1990).

Example 2. *Calculation of Percent Agreement*

		Technician 2	
		Positive	Negative
Technician 1	Positive	a	b
	Negative	c	d
	Percent agreement = $(a + d)/(a + b + c + d) \times 100$		
	Percent positive agreement = $a/(a + b + c) \times 100$		

Neither overall nor percent positive agreement takes into account the fact that some amount of agreement between two observers will be due to chance alone. The extent of agreement between two readers beyond that due to chance alone can be estimated by the kappa statistic (Example 3) (Agresti 1990; Fleiss 1981; Landis and Koch 1977). In comparisons of more than two categories, a weighted kappa approach allows consideration of the fact that disagreement between some categories may be more serious than disagreement between other categories (Cohen 1968). Like the sensitivity and specificity, variations of the kappa statistic are limited by the fact that most underlying traits are not dichotomous, and different cut-off levels can affect the value of kappa (Maclure and Willett 1987). Interpretation of the kappa statistic should also take into account the fact that

kappa can be affected by the prevalence of the condition: for a fixed sensitivity and specificity, kappa tends towards zero as the prevalence of the condition approaches either zero or one (Thompson and Walter 1988). Additionally, high values of kappa can be obtained if the marginal totals of the contingency table are not balanced (Feinstein and Cicchetti 1990; Maclure and Willett 1987; Thompson and Walter 1988).

Example 3. *Calculation of Kappa for a binary measurement variable*

		Technician 2		
		Positive	Negative	**Totals by Technician 1**
Technician 1	Positive	45	5	50 (61.0%)
	Negative	2	30	32 (39.0%)
Totals by Technician 2		47 (57.3%)	35 (42.7%)	82 (100%)

$$\text{Kappa} = \frac{(\text{Proportion Observed Agreement} - \text{Proportion Expected Agreement due to Chance})}{(1.0 - \text{Proportion Expected Agreement due to Chance})}$$

Proportion Observed Agreement, $P_o = (45 + 30)/(45 + 5 + 2 + 30) = 0.91$

Proportion Expected Agreement due to chance, $P_e = \frac{(50\times 47)/82 + (32\times 35)/82}{82} = 0.52$

$$\text{Kappa} = \frac{P_o - P_e}{(1.0 - P_e)} = \frac{(0.91 - 0.52)}{(1.0 - 0.52)} = 0.81$$

Common measures of agreement used to assess reliability for continuous measurements (such as blood pressure readings) are the *correlation coefficient*, the *intra-class coefficient,* the *average error*, and the *coefficient of variation*. *Linear regression* techniques can also be used to check for systematic differences (cf. Chap. II.3 of this handbook).

Although the *Pearson's correlation coefficient*, *r*, is one of the most frequently used measures of agreement in the medical literature, its use is often not appropriate (Altman and Bland 1983; Szklo and Nieto 2000). For one, the correlation coefficient is equally high when both observers read the exact same value, and when a systematic difference (bias) exists between observers but the readings vary simultaneously. The value of *r* is also very sensitive to extreme values and the range of values, with a broader distribution of values yielding a higher *r*. While the Spearman correlation coefficient r_s may be more appropriate to assess the comparability of the rankings of readings, and would moreover be less sensitive to outliers, it does not address the main problem of the inability to detect systemic differences between observers.

The *intra-class coefficient* (ICC), or the reliability coefficient, estimates the proportion of the total measurement variability due to the variation between individuals (Fleiss 1981). The ICC is analogous to the kappa statistic used for categorical variables, and the value can be interpreted in a similar manner. The ICC is a true measure of agreement in that it combines information on both the correlation, and the systemic differences between readings (Deyo et al. 1991). As with the cor-

relation coefficient, however, the ICC is affected by the range of values in the study population.

Other commonly used measures of variability are the *average error*, and the *coefficient of variation* (CV). The average error is the ratio of the mean absolute difference of pairs of measurements to the overall mean value of the measurements. The coefficient of variation is the standard deviation expressed as a percentage of the mean value of sets of replicate observations. In a reliability assessment, the CV would be calculated for each pair of observations, and then averaged over all pairs of original and replicate measures. A limitation of the CV and average error is that both measures may reflect the magnitude of the mean value more than the magnitude of the measurement error (Canner et al. 1983). An alternative measure that has been suggested for assessing variability is the increase in the among-participant standard deviation, the I_{APSD} (Canner et al. 1991). This measure can directly determine the impact of measurement error on the overall among-participant variability for a variable of interest.

Linear regression techniques can estimate systematic differences between readers which are reflected in the slope and intercept of the regression model. One drawback of using regression to assess reliability, however, is that measurement error occurs in both the dependent and independent variables, violating the assumption of an error-free independent variable required for regression (Altman and Bland 1983). However, only under unusual circumstances would measurement error lead to confusing or uninformative results.

Planning Data Management

13.3.7

The management of data in a large epidemiological study can be a formidable task. The sheer volume of data for a sizeable study with extended follow-up can become quite overwhelming, as illustrated by the following example. If 100 data elements are to be collected for each participant in a cohort study with 100,000 participants, the data collected at the end of each data collection cycle are comprised of 10^6 distinct data elements. Let us say that in order to update exposure and outcome information, data are to be collected yearly for each participant for ten years. This increases the amount of data being collected by an order of magnitude, to 10^7 distinct pieces of data. Superimposing on this volume of data the errors that can occur during data recording, transcription, and transfer of data to an electronic medium, it is easy to see how data quality can be compromised without careful planning of how data are to be managed. The potential magnitude of the task of data correction is also clear.

The first step in planning a data management system is to define what data will be collected and how often data will be collected, keeping in mind that as the volume of data increases, ensuring data accuracy becomes more difficult. In order to further minimize the amount of unnecessary data collected, the chosen data variables should be prioritized, and a "tolerance" of error established for each data field. For example, it might be decided that all values of crucial

Table 13.2. Common Measures of Agreement, interpreted in the context of two separate technicians reading the same data

Statistic	Range	Type of Data	Interpretation
Sensitivity and Specificity	0 to 100%	Categorical	Sensitivity: ability to correctly identify individuals who have the disease or exposure characteristic of interest. Specificity: the ability to correctly identify individuals who do not have the disease or exposure characteristic of interest.
Overall percent agreement	0 to 100%	Paired categorical variables	The proportion of all readings that are categorized in the same way by two different observers. Higher value means better agreement.
Percent positive agreement	0 to 100%	Paired categorical variables	The proportion of all non-negative readings that are categorized in the same way by two different observers.
Kappa statistic, κ	−1 to 1 (rarely below 0)	Paired categorical variables	The extent of agreement between two readers beyond that due to chance alone. Higher value of kappa means better agreement.
Weighted Kappa	−1 to 1 (rarely below 0)	Paired categorical variables	The extent of agreement between two readers beyond that due to chance alone, allowing for consideration of partial agreement.
Pearson's correlation , r	−1 to 1	Continuous ordinal variables	The degree to which a set of paired observations in a scatter diagram approaches the situation in which every point falls exactly on a straight line. −1 is perfect negative correlation, 1 is perfect positive correlation.

Table 13.2. (continued)

Statistic	Range	Type of Data	Interpretation
Spearman's correlation, r_s	−1 to 1	Non-parametric ordinal variables	The degree to which ranking of measurements is consistent between two readers.
Intraclass Correlation Coefficient, ICC	−1 to 1 (rarely below 0)	Continuous variables	The proportion of the total measurement variability due to variation among individuals. Analogous to the kappa statistic, but for continuous variables. Higher ICC means better agreement.
Linear regression, β, c		Continuous variables	Yields a measure of the intercept c and slope β of the regression function.
Coefficient of variability, CV	0 to 100%	Continuous variables	The standard deviation expressed as a percentage of the mean value of two sets of paired observations. Lower CV means better agreement.
Average error	0 to 100%	Continuous variables	The ratio of the mean absolute difference of pairs of measurements to the overall mean value of the measurements. Lower average error means better agreement.
I_{APSD}	0 to ∞	Continuous variables	The percentage increase in among-participant standard deviation due to intra-observer measurement error.

data variables (e.g. disease outcome) should be checked against written questionnaires, but auditing a random sample of questionnaires is sufficient for other fields.

The next key step in data management planning is to define essential identifying information for the study data. *Identifiers*, generally known as *key*, *header*, or *ID* fields, are fields that allow each form to be uniquely identified and correctly related to other forms (Hosking et al. 1995; Hosking and Rochon 1982). Study identifiers are usually located in a standard header section of the form. Entry of an incorrect number into one of these fields can cause the entire data record to be processed incorrectly.

Most studies require at least four types of identifiers: study identifiers, participant identifiers, form-type identifiers, and time-point identifiers. Depending on the study, other identifiers (e.g., family identifiers) might also be necessary. *Study identifiers* designate the sponsor, study, protocol, or sub-study. *Participant identifiers* uniquely identify the study participant. In general, a study-created participant identifier is preferable to a natural identifier such as participant name or social security number, especially in a climate increasingly concerned with participant confidentiality. Encode information about participant characteristics (such as a field site code) into the participant identifier is useful, since this allows later classification of participants by their identifiers alone. *Form identifiers* identify a particular questionnaire, and often take the format of a two- or three-character abbreviation. Form identifiers in longitudinal studies should be planned so that multiple versions of each form can be accommodated. Adding a −1 at the end of the form identifier, for example, allows for future versions to end with the suffix −2 or higher. *Item identifiers* are assigned to each question on a form. While item identifiers bearing a one-to-one relationship to database fields might be useful for data analysis, this can become confusing if the study forms or database are revised. Data management systems that track the relationship between each database field and a corresponding item number in each form version provide a useful alternative.

Once data identifiers have been selected, general data management considerations that need to be addressed including identifying how data will be entered (electronically versus manually), who will do the data entry, what software will be used, what types of edits will occur during data entry, how queries will be generated, communicated and resolved, how suspicious values will be treated, and how corrections will be implemented and documented. The remainder of this section will be devoted to these considerations.

Design of a Data Management System

In a multi-center study, data are typically collected at various field centers and then sent to a coordinating center for processing, storage and analysis. Table 13.3 provides an overview of the steps involved in data management. While newer approaches to data management exist (e.g., web-based systems), these approaches rely heavily on specialized automated systems for data collection, entry and audit-

ing. Since logistical barriers of cost, lack of expertise and low computer literacy currently render these systems impractical for many investigators, this chapter focuses on a more traditional approach to data management. Many of the underlying principles remain relevant to the newer data management approaches, which are addressed at the end of the section.

Table 13.3. Overview of the Traditional Data Management Process*

Steps in data processing	
Data collection and mailing	Complete forms at clinic/in field. Visually review form while participant is in clinic (visual editing). Mail original copy to coordinating center, keep copy at clinic. Create standard packing list for mailing
Receipt and conversion to electronic format	Receive forms at coordinating center. Acknowledge receipt of forms from clinic using postcard or electronic mail. Log receipt of forms into computer (including form number, ID code, date completed, date received, and unique log number). Key the form. Verify keying.
Forms processing, posting and backup	Process form through an edit program that checks type and range of each field, as well as internal consistency of form. Generate computer edit report. Check edit report and initiate appropriate error correction procedures. Back-up edited forms. Post forms to master file. Back up master file. File form.
Clearance and archiving	Run further checks on data to ensure that posted data are consistent with other data on file. Review edit reports that result from checks and initiate appropriate error correction procedures. Document master file contents and prepare file for archiving.

* adapted from DuChene et al. (1986)

Data Recording and Visual Editing

The measurement and recording of data from study participants usually occurs at the field center, and initial data checks are generally conducted by field staff personnel. Field center interviewers or technicians should check data for consis-

tency as it is being collected, while the study participant is available to clarify any immediate discrepancies, errors, or out-of-range characteristics.

For technical measurements, an independent review of samples by two or more readers should be performed on all, or a subset of samples. This allows later assessment of validity, and enables investigators to track down sources of error. On completion of the data form, field center staff should perform a routine review of forms to establish that the questionnaire is complete, that skip patterns have been followed, and that the data values appear reasonable. If routine review of the form does not identify any unusual data, the form can be processed further. Including an indication of who reviewed the form will facilitate later examination of the editing process.

Data Entry

Almost universally, epidemiological data are entered into electronic databases for storage and analysis. The processing, storage, and analysis of study data usually occurs at the data coordinating center. Errors that can occur during the processing and storage of data include keying errors, inaccurate data transcription, and programming errors (Arts et al. 2002). In the Hypertension Prevention Trial, key error was found to be the major source of data entry error, with 5.2/1000 errors out of an overall error rate of 6.9 errors per 1000 data items being key errors (Prud'homme et al. 1989).

Most automated data entry systems allow a variety of mechanisms for checking data. As data entry is initiated, form identifiers are checked for validity and consistency. *Range checks* during data entry can be used to electronically limit the data type, or the range of possible values at entry. For example, date fields can be programmed to accept only valid dates, or table look-up systems can restrict the values of categorical data to a limited number of possible values. For continuous data, many studies use normal population ranges of a variable to flag outliers. While programmed range checks are a useful tool, retaining some flexibility to correct errors at the time of data entry is important, since too many restrictions on modifying data at entry can lead to a higher error rate (Crombie and Irving 1986).

Data accuracy can also be improved by the use of *double data-entry*. The independent keying of data twice, however, does not prevent all types of error. Examples of errors that would not be reduced by the double entry of data include errors in transcription, or misinterpretation of data in the same way by two data-entry operators.

If the data are to be manually entered, personnel should be masked regarding exposure or outcome status (depending on the study design), to prevent the possibility of observer bias. Additionally, the electronic database should have a provision to indicate who entered the data to allow for later review of data-entry performance.

The use of electronic technology for data entry as an alternative to manual data entry is gaining in popularity. Software is available for scanning forms directly into an electronic database using optical character recognition (OCR). The accuracy of scanning software is quite variable, however, and a process to check scanned

data should be in place. In general, OCR is better suited to numeric or check-box responses than to hand-printed characters. Another method of electronic data entry is computer-assisted data collection (CADC), whereby interviewers directly enter participant responses into a computer file. This technology completely circumvents the need for transferring data from paper to an electronic medium, thus eliminating errors associated with this process. A CADC system can automatically enforce skip rules, require completion of required fields, and flag suspicious values for correction while the study subject is still present. Since errors in CADC data cannot be compared later with a paper form, however, these systems need to include as many ways of checking data accuracy at entry as possible. One way to allow for examination of inconsistent values is to tape record interviews while data collection is occurring.

In a pilot study of CADC, five study staff members with no prior experience using a CADC system were trained and asked to administer both CADC and paper-based interviewers to sixteen study participants. All five staff members preferred the CADC system, indicating faster and more accurate data entry and less likelihood of erroneously skipping an item. Ten of the sixteen pilot study participants had no preference between paper and CADC, and six preferred CADC. Although the median time for data collection at the reception, examination and interview stations was slightly longer for CADC than for paper interviews, the CADC data are already partially edited and in machine readable format, whereas data from the paper forms still had to be edited and keyed. The percentage of suspicious data values was similar for each method, but 21 of the 25 suspicious data values were identified and corrected at the time of collection using CADC, compared to 1 out of 23 suspicious values corrected with the paper system (Christiansen et al. 1990).

Other recent methods of data collection for epidemiological studies include the use of electronic-mail ("e-mail") (Kiesler and Sproull 1986; Paolo et al. 2000) or internet-based surveys (Baer et al. 2002; Blackmore et al. 2003; Rhodes et al. 2003; Silver et al. 2002; Turpin et al. 2003). E-mail questionnaires have been reported to have a faster rate of return and more thorough completion of returned questionnaires, but response rates have generally been lower than for mail questionnaires.

The basic process for internet-based, or web-based, data collection is the translation of the study questionnaire into an internet language (HTML, or hypertext markup language) and posting of the questionnaire onto the World Wide Web. Respondents then complete the survey using a point and click interface. The survey is generally visually and functionally similar to traditional surveys.

Web-based data collection provides several advantages over paper form data collection. For one, researchers can reach populations that previously might have been inaccesible due to geographical or cultural boundaries. Use of the web may also speed up the time of data collection, since no testing site or appointment scheduling is necessary, and the need for data entry by study personnel is eliminated. Web-based systems can also minimize the variation due to differences in survey administration, interviewer interpretation and entry of data. Since complicated branch and skip patterns can be programmed into the survey, the amount of interviewer or respondent attention necessary is reduced. Costs can drop dramat-

ically with the use of web-based data collection, as there is no need for printing, mailing, and data collection personnel. Web-based surveys also provide a greater degree of anonymity for the collection of sensitive personal information (Baer et al. 2002).

It is important to realize, however, that depending on the situation, some advantages of web-based systems can become disadvantages. In certain populations or countries, for example, the cost of printing, mailing and administering a paper questionnaire might be considerably less than the cost of setting up a web-based system and providing training and access to study participants. Other disadvantages of web-based data collection include the possibility of selection bias when choosing a study population, and security problems during data transmission. The issue of computer users being unrepresentative of the general population can be overcome to some extent by providing internet access to a randomly sampled study population (Silver et al. 2002). Literacy or language barriers, however, may still prove to be an issue.

Incorporating strict security measures in an electronic data entry system is crucial to maintaining data confidentiality, and can require considerable time and monetary resources. In some instances, it might be possible to provide a quick solution to this problem by linking the survey security to an existing high-security system, such as a university network.

While the use of web-based systems is a promising avenue of data collection for studies, such systems require considerable expertise for adequate set-up. Often, initial versions of web-based questionnaires present frustrating technical problems to users, and may require several iterations before a working system is in place. Web-based systems may be inappropriate for populations that are not computer-literate. It may be more difficult to adress ethical concerns which arise during the course of a study in the context of web-based data collection. For example, the investigators still bear responsibility for verifying informed consent, or for providing local targeted support in case the respondent needs a referral as a result of the research. Additionally, data entry errors by users can still occur. For these reasons, the pre-testing of data instruments, post-entry error-checking, and other forms of data quality control described in this chapter are as crucial for more technologically advanced data collection as they are for more traditional forms of data collection.

Data Audits

Once data have been entered, they must be submitted to further accuracy checks. One method of assessing data accuracy is to perform a series of consistency checks, such as ensuring that the date of birth and age of a participant are in agreement. Reviewing samples that fall outside some number of standard deviations of the mean is a sensible alternative way to check data. More formal statistical methods for detecting outliers can also be used (Barnett and Lewis 1994; Vardeman and Jobe 1999). The importance of using range checks is illustrated by a simulation in which different rates of entry error were introduced into a constructed dataset,

and simple range checks were used to identify and correct outliers. Even with a random entry error rate as high as 20%, population means remained very similar after the correction of unusual values, regardless of study sample size (Day et al. 1998). Error rates similar to those achieved with double data entry were achievable when extensive logic checking of fields was incorporated (Mullooly 1990; Neaton et al. 1990).

In instances where an unusual value is detected, a data quality query should be generated either manually or automatically. A system for reporting and responding to such queries needs be conceptualized during study planning, along with the designation of individuals responsible for checking and responding to questions. The automatic generation of regular quality control reports including summary statistics such as the number of queries by form and data field, or the percentage of error-free forms, can aid the systematic processing of data. Section 13.4 of this chapter addresses the processing and resolution of error queries in more detail.

Comparing the number of forms that are edited using automated checks at the data processing center to the number of forms recorded in the batch sent by the field center allows the identification of forms that are lost during keying. Additionally, a random sample of data forms should be compared to the electronic data submitted to check accuracy of data entry.

Once routine edits have been completed, the data form can be posted directly to a *master file* for smaller studies, or to a *distributor file*, for larger studies. In the Multiple Risk Factor Intervention Trial (Mr Fit), the edited form was transferred to a distributor file, which held all the forms that were edited in a day. At the end of the day, forms held in the distributor file were transferred to one or more *transaction files*, which served as temporary storage until the next scheduled update of the master file. The use of transaction files allowed investigators the flexibility to resolve discrepancies before the data were added to the master file. Transaction files were generally copied to daily backup tapes so that data could be retrieved to the time of the last back-up in case of processing errors, machine failure, or other accidents (DuChene et al. 1986).

Forms Posting

In general, it is best to keep the interval between data collection and entry as short as possible. If it is possible to process forms as they are generated, this is preferable (Meinert and Tonascia 1986). However, if batch processing is found to be more convenient, the scheduled time between subsequent postings of information from the study transaction files (raw data) to the master file should not be longer than two weeks. During forms posting, data fields from the transaction files are copied to the location in the master file(s) specified in the *data dictionary* (a database of information used to edit, document and control the processing of forms through the computer system). The data management system should be programmed to reject the form if errors are detected in the data identifiers, or if data are found to already exist in the master file (unless the form to be entered is a correction form). Personnel at the data coordinating center can then review and resolve discrepancies

in rejected forms. If fields need to be modified in the master file, changes should be explained and documented in the electronic file as well as on paper.

Backup of Raw Data

Once the forms have been posted to the master file, all transaction files containing the posted forms should be copied on to a tape or other electronic medium such as compact disc (CD) or digital video display (DVD), and stored offsite. In the event of a major system failure or destruction of the master file (in a building fire, for instance), the offsite copy will allow recreation of the master file.

Clearance

After the data are posted to a master file, computer edits of the master file allow consistency checks between fields on different forms. For example, an individual's height should remain constant over forms. It is informative to flag inadmissible values, as well as unlikely values. Additional within form checks can also be performed at this time.

Archiving

When within-form verification and across-form clearance are complete, and data on the master file are finalized, the master file should be copied on to at least two tapes and stored off-site. These tapes should be read regularly to check for deterioration. If a back-up tape cannot be read, a new copy should be made.

13.3.8 Quality Assurance Committee

The most carefully designed quality assurance program cannot function efficiently without the assignation of responsibilities for various quality monitoring tasks to specific individuals, and the existence of effective communication channels between study personnel. In many large studies, a quality assurance committee is formed to oversee the quality of data collection (Knatterud et al. 1998; The Chemical Manufacturers Association's Epidemiology Task Force 1991; US Environmental Protection Agency (EPA) 1989). The quality assurance committee addresses quality issues throughout the life of the study, from protocol development to the responsible archiving of data. The quality assurance committee is also responsible for reviewing study compliance with written quality assurance/control procedures, and for evaluating interim analyses. For large studies, a data monitoring committee made up of external quality assurance auditors supportive of the protocol objectives and study design might be warranted (Fleming 1993).

13.3.9 Communications

The effective resolution of study quality issues is highly dependent on the quality of communications between study personnel. Many of the quality assurance mechanisms already described in the chapter *contribute* directly to improved communication. Examples include the training of personnel, and the definition of standard

operating procedures. Other quality assurance mechanisms *depend* critically on communication for their implementation. In order for queries to be resolved effectively, study personnel need to know who to submit queries to, and how these queries should be submitted. Structures for transmitting resolved data queries back to data entry personnel are also needed. The scheduling of regular meetings between study personnel is crucial for maintaining study communications. Emphasizing the rationale for quality control and the need for wholehearted support for quality control measures is important, since quality control measures will fail if they are perceived as nit-picky and burdensome (Cooper 1986). One or more individuals should be designated responsible for preparing and disseminating the minutes of study meetings. More generally, communication structures should be in place to communicate the intent, conduct, results and interpretations of the study to study personnel, study participants, and the scientific community. In certain situations, other parties that might need to be informed of study results include health care providers, policy makers, or the media.

Cost of Quality Assurance

13.3.10

Clearly, the implementation of quality assurance and quality control measures add to the cost of a study. While some expenses, such as the cost of routine data editing, or the re-checking of statistical analyses may be impossible to estimate, cost information can be projected for other aspects of quality assurance, such as training, site visits, and external quality control programs (Knatterud et al. 1998). Considering the cost of various quality control measures early in the planning process allows for development of a realistic and feasible program that is more likely to be executed. Priorities for data quality should be set at this time. While certain aspects of data quality should not be sacrificed regardless of the expense, a compromise might be possible in other instances. For example, a costly, time-consuming measure of exposure might be collected for a sub-sample of study participants and this information can be used to validate a cheaper exposure measurement used for all study participants.

Ethical Considerations

13.3.11

Ethical considerations are perhaps the most important set of considerations in a study (for a general discussion see Chap. IV.7 of this handbook). Epidemiological research should never lose sight of the fact that data are derived from human beings. Studies such as the Tuskegee Syphilis Trial (US Department of Health Education and Welfare (DHEW) 1973) which followed the progress of untreated syphilis in black men even after effective treatment was available may now seem shocking, but it is well to keep in mind that throughout most of the trial, the investigators did not find their research particularly objectionable. The thorough consideration of ethical issues raised by a study (mandated by law in most countries) will hopefully prevent a future generation of scientists from looking back at present-day trials with regret.

The human subjects section of the protocol must describe whether the study protocol imposes any physical or psychological risk to the participants. Potential benefits of the study should also be noted, with an explanation of whether benefits will be accrued by study participants themselves, or whether the study is expected to benefit others in the future. The cost-to-benefit ratio should be weighed and discussed. Studies that involve primary data collection generally need to obtain informed consent from study participants. Consent forms should include, at a minimum: contact information for personnel available to answer questions about the research; the purpose of the study; eligibility requirements; the expected duration of participation; possible harm that the subjects could incur; expected benefits to subjects or to others; information on the voluntary nature of participation, and a statement indicating the right to withdraw from the study at any time (The Chemical Manufacturers Association's Epidemiology Task Force 1991). The study eligibility criteria are also subject to ethical considerations, both in terms of inclusions (different racial/ethnic groups and both genders should be adequately represented) and exclusions (special justification is needed for study of vulnerable groups, such as pregnant women, children, or incarcerated individuals). Adequate provisions for maintaining data confidentiality and the privacy of individuals should be described. For example, investigators might plan to store hard copies of sensitive data in locked cabinets with limited access and remove personal identifiers from datasets used for analysis. Automated data management systems should have password control, users should be logged out after a period of inactivity, and the copying of data should be discouraged (Wyatt 1995).

13.4 Quality Considerations During Study Conduct

Before data collection is initiated, all data collection procedures should be reviewed and approved by the lead investigators. Data forms and equipment should have been tested, and certified ready for use.

If rigorous quality assurance procedures have been planned prior to study initiation, quality control activities during study conduct mainly consist of the implementation of these procedures. The study protocol should be followed, personnel should be trained according to established standard procedures, and data collection should proceed with all quality assurances in place. Any deviation from standard operating procedure should be authorized by the Steering Committee.

The importance of periodic examination of data by study investigators, data coordinators, and data entry personnel while the data are being collected cannot be overstated. Examination of data trends by center, over time, or by technician (for example), can identify flaws in data collection early on. Even simple plots and graphs of data can identify sources of error. When data errors are identified, steps should be taken to correct the data in a timely manner. In some cases, statistical

adjustment can be used to correct data drift. When this is not possible, data might have to be thrown out, or completely reprocessed. In order to generate a written audit trail of data, any changes made in the data should be documented.

Training and Certification 13.4.1

The importance of training and certifying all study personnel has already been underlined in Sect. 13.3.3. While many study investigators are aware of the need for standardized operating procedures, information regarding these procedures is often lacking in study descriptions. While 244 original research articles in three emergency medical journals (1989–1993) described data collection by means of chart review, only 18% mentioned training of abstractors, and periodic abstracter monitoring was reported in a mere 3% of these articles (Gilbert et al. 1996).

Detailed practical guidelines for training and quality control management for study interviewers, data abstractors, and biomedical technicians are available in the literature (Edwards et al. 1994; Fowler and Mangione 1986, 1990; Reisch et al. 2003). This section summarizes some of the main considerations.

Training

Training procedures should ideally involve all staff and procedures. While centralized training of all study personnel might be desirable in terms of increasing the comparability of data collection between sites and allowing study personnel from different sites to interact with each other, the expense of bringing personnel to a central training site for all their training can be considerable. Additionally, site-specific questions might arise that cannot be adequately addressed during centralized training. An optimum strategy might be to use both types of training. Table 13.4 provides an overview of the training process.

Certification

Following initial training, study personnel should be certified to perform specific procedures. Regular re-training is desirable to prevent data drift. Re-training might also be necessary if a specific study technician is found to be introducing a systematic error into the data, or if the study protocol changes. Any re-training should be accompanied by recertification.

While the interval between re-training and certification varies from study to study, the Atherosclerosis Risk in Communities study (ARIC) used a 90-day interval, since a six-month interval was found to allow too much drift to recognize and correct digit preference. More timely feedback was also needed in the Cardiovascular Health Study (CHS) (Hill 2003).

Maintenance and Calibration of Equipment 13.4.2

Study equipment should be inspected and calibrated at regular intervals in accordance with the study protocol. In the event of equipment breakdown, equipment

Table 13.4. Overview of Training*

Steps in data processing	
Training manual	Educational training manual is sent to all sites for review. The training manual consists of some or all of the following: a study overview, information on the relevant procedure, quality assessment procedures, data forms with instructions (e.g. for abstraction or interview), quick reference sheet for all variables, glossary of terms, standardized training examples, and relevant articles from the literature.
Standardized training examples	Training examples should be prepared for key study variables. For instance, study personnel might be asked to note blood pressure measurements from a training tape.
Individual orientation	Two or more individual orientation sessions should be arranged with the onsite data collection team, and with the lead study co-ordinator and/or study investigator. Additional sessions can be scheduled at the discretion of the site co-ordinators.
Double-review of initial data	The first few examples of data collected (by chart abstraction, interview, or a biomedical procedure) should be repeated by a more experienced member of the data collection team. Discrepancies can then be reviewed. Queries should be entered into an audit form and sent to the lead study co-ordinator to assist with later tracking of problematic data.
Regular double-review	Performing regular double review for a small sample of data (e.g. once a month) can prevent data drift over time. Review of data at a later time is facilitated by audio or video taping of interviews or biomedical procedures.
Regular conference calls/meetings of field staff	Regular study conference calls can include a training component if examples of data collection problems are brought up for discussion during each call. An updated decision log containing a summary of discussions held and decisions made during these conference calls can be distributed among study personnel.
Regular site visits	Review of data collection procedures during site visits by the lead study coordinator and/or lead investigator.
Retraining	Retraining of study personnel might be necessary if substantial time has passed since initial training, a systematic bias in data is detected, or the study protocol changes.

* adapted from Reisch et al. (2003)

may need to be replaced. If the new equipment is similar to the equipment already being used, then calibration before use is sufficient. When replacement of existing equipment is desirable because a new model or instrument is more accurate or efficient than the existing equipment, data should be collected using both the old and the new instrument for a defined period of time, so that comparability of measurements can be established.

Implementing Data Management 13.4.3

The data management process has already been described in detail in Sect. 13.3.7. During study conduct, the planned data management system is implemented, and refined as necessary.

Tracking and Monitoring of Data

The effective tracking and monitoring of data as data collection is in progress is essential to the timely detection and correction of errors. Monitoring should occur for subject accrual, data acquisition, and data quality. Automated tracking systems can greatly assist this process, and have been used successfully in epidemiological studies as early as 1981 (McQuade et al. 1983). Data that are collected by hand should be recorded directly, promptly, and legibly in ink. Four different types of monitoring are recommended: *pro-active efforts* to improve data, *observation of data collection*, *review of computer-generated checks and summary reports*, and *examination of data*.

When possible, data quality should be improved by *pro-active efforts*. Automated reminders of when patients are due for study visits for time-dependent variables (e.g. levels of an exposure biomarker) can prevent the collection of data that is later deemed of poor quality or unusable. Target dates for follow-up visits can be defined by the participant's entry date rather than the date of the last visit, in order to prevent scheduling deviations from carrying over to future visits.

Direct or indirect *observation of data collection* can also identify errors in a timely manner. An unobtrusive way to monitor interviewers for delivery and adherence to protocol is to audio-tape interviews. Measurement techniques for biomedical or laboratory technicians can either be videotaped, or directly observed by senior technicians or other qualified study personnel.

Regular *review of computer-generated queries and summary reports* of data quality can alert the investigators to a variety of data errors, including participant ineligibility, data outside the expected range, and variation in data quality by data field, site, or technician. Active examination of data during collection is crucial. Summary statistics and plots of data by technician, site or time can identify unusual trends. For example, an examination of data from the Hypertension Prevention Trial revealed that nearly 29% of the baseline systolic blood pressure readings from one clinic ended in the digit 2. This could be traced to measurements made by one technician, who recorded a number ending in the digit 2 for over 60% measurements (Canner et al. 1991). When a data collection flaw is identified,

further error should be prevented by tracking down the source of the problem, and taking corrective action.

Keying errors may be identified by periodic audits of the database against source documents. Rather than check all the data, a random sample of data fields can be selected to check for keying errors. When creating the test sample, it is important to ensure that a broad cross-section of data is included (for example, both numerical and character fields should be checked). One method for sampling a variety of fields is to choose a random sub-sample of forms, and look at all fields within those forms.

Corrective Actions

Moving back to the datascope for a moment, we recall that the identification of data errors is only the first step in data quality management. In order to reach the ultimate goal of valid data, these errors need to be corrected. The process for revising data should be as systemized and well documented as the process for locating errors. While the routine correction of careless mistakes while data entry is in progress need not be reported, data errors that are identified after initial data entry should not be changed by data entry staff until the query has been checked. A paper trail should be initiated for each problem, with the initial query describing the problem, and the date it was detected (Fig. 13.2). The individual(s) responsible for query resolution should then investigate the query, and provide a response explaining why the problem occurred. Finally, the query documentation should indicate how and when the problem was resolved. If data from a form are found to be incorrect, they should be identified as incorrect rather than erased, and the correct values should be inserted (Knatterud et al. 1998). In some cases, unusual values will be confirmed to be correct, in which case they should be retained in the database with documentation.

Occasionally, errors identified during study conduct may lead to changes in the survey instrument or other study equipment. In such cases, it is crucial that the version of the form or equipment used to collect data is recorded in the database. If a new data check is added, either as a result of a query or as an additional precaution, old values in the database should be edited using the new rules in order to keep data consistent.

Tracking the time taken for corrective actions allows areas of delay to be identified and resolved for future queries. In most longitudinal studies, data are analyzed while data collection is still in progress. In such instances, one might want to exclude data that are under query from the master database until the problem is resolved. The inclusion of a "status" field for data would allow investigators to check whether values were acceptable or unacceptable (Gassman et al. 1995).

13.4.4 Site Visits

For multi-center studies, site visits to observe operations allow greater understanding of site-specific data collection issues, and provide an opportunity to recognize and correct faulty systems (Gassman et al. 1995; Knatterud et al. 1998; Prud'homme

Query

Subject ID: 111770
Form: 121 **Item Questioned:** 5, 6a
Date of Visit: 08/28/02 **Visit Number:** 2
Description: Subject claims to be a former smoker (ev_smok = 2), but reports currently smoking five cigarettes a day (cur_cig = 5).

Date: 12/6/02
Initials: PR

Response

Form to correct: 121 **Item to correct:** 6a
Old value: 5
Correct value: 0

Explanation: Checked subject's medical record and past questionnaire. Subject is a former smoker.

Date: 12/11/02
Initials: DR

Documentation

Correction: Value of cur_cig has been changed from 5 to 0.

Date: 12/20/02
Initials: TN

Figure 13.2. Example of a Data Query Form

et al. 1989). Scheduling a site visit is recommended shortly after initiation of patient recruitment, and when the data collection at the site is drawing to a close. Additional site visits should be scheduled for long-term studies.

The size of the site visit team can vary, and is dictated by the nature and purpose of the visit. A typical site visit team might include the study principal investigator (or representative), the director of another field site, the data

coordinating center director, the study project officer, and selected resource personnel. During the site visit, the site visit team would meet with the director and staff of the unit, and hold private conversations with key support personnel. The site visit should include a thorough review of staffing requirements, recruiting, training and certification, and communication structures. Site visitors also have a chance to observe data collection, check data management, and review data quality monitoring. Specific activities might include observation of whether field technicians follow the study protocol, inspection of study records and documents storage, and review of the operation and maintenance of local data systems. Following the site visit, the leader of the site visit team should prepare a written report of the visit based on input from the entire team. The site visit report should describe any systematic errors that were identified in data collection, and provide recommendations on how to rectify the situation. A formal response to the report should be prepared by the staff at the study site.

13.5 Quality Considerations After Data Collection

Once data collection for the study is complete, the task of analyzing and interpreting the data begins. The study investigator should yet again consult the datascope to check for possible biases and errors that need to be resolved in order to form a clear picture of the relationship under study.

13.5.1 Reporting Response Rate

If individuals who agreed to participate in the study were different in some important way from non-respondents, the study results could well be biased. For example, non-respondents to questionnaires might be of poorer health or more likely to be smokers than respondents (Shahar et al. 1996). Studies that have followed respondents and non-respondents to questionnaires have reported that non-respondents have a significantly higher risk of myocardial infarction, cancer mortality, and all-cause mortality (Bisgard et al. 1994; Heilbrun et al. 1991).

Calculating the study *response rate* gives a first indication of whether the investigator should be concerned about possible bias in the results. Generally, the higher the study response rate, the less need to worry about selection bias affecting the results. The simplest approach to response rate calculation is to divide the number of surveys received by the number of surveys sent. However, this does not account for factors that can affect the response rate such as undelivered questionnaires, ineligibility of subjects who completed questionnaires, or substitution of the intended recipient with another subject. Typically, the numerator and denominator of the response rate are adjusted to reflect such factors. Standard definitions and

methods to calculate survey response rates are provided by the American Association for Public Opinion Research (2000), or the Council of American Survey Research Organizations (CASRO).

For cohort studies, the simplest way to estimate the *follow-up rate* is to the divide the number of participants seen at the last visit by the number of participants initially enrolled. Again, different assumptions about individuals lost to follow-up yield different numbers for the follow-up rate.

Since different methods of calculating the response rate might be appropriate for different studies, the choice of the response rate formula is less critical than the identification and reporting of all the elements that enter the calculation (Table 13.5).

In general, response rates to questionnaires have been decreasing in the United States, and perhaps elsewhere (Kessler et al. 1995; Steeh 1981). Data from a nationwide survey in the United States (the Behavioral Risk Factor Surveillance System, BRFSS) indicate that response rates from random digit dialing have declined from a median of 68.4% in 1995 to a median of 55.2% in 1999 (Centers for Disease Control and Prevention (CDC) 1999). A review of 82 case-control studies published in the *American Journal of Epidemiology* (1988–1990), *Epidemiology* (1997–1999) and *Cancer Epidemiology, Biomarkers and Prevention* (1997–1999) reported a 0.2% and 0.44% decrease in reported response per year for cases and controls, respectively (Olson 2001). The same article reported an average response rate of 76.1% for cases and 71.5% for controls. A review of 321 distinct mail surveys published in a broader spectrum of United States journals in 1991 reported an average survey response rate of 62% (Asch et al. 1997).

Regardless of the exact value of the response rate, the characterization of non-respondents is crucial in order to assess whether a bias is present, and if it is, how the results of the analysis might be affected. Clearly, describing the non-respondents becomes more important when a study has a low response rate. Whenever possible, a brief survey should be administered to non-respondents to collect limited data for comparison with respondents. Otherwise, assessing available data on demographics, exposure or outcome will allow some assessment of possible bias.

Analysis 13.5.2

Before proceeding to analysis, the study data should be tested rigorously to check for *residual errors* that remain after all data processing and routine quality assurance activities are complete. Range checks provide one way to examine whether the data seem reasonable. Simple queries such as checking that the recorded age in years is consistent with the date of interview minus the recorded date of birth, can also help to detect errors.

Once the investigator feels confident that there are no obvious flaws in the data, the next step is to understand the data by conducting exploratory data analysis using univariate and bivariate summaries, as well as plots and graphs of the data. More complex exploratory analysis of the data should be guided by

Table 13.5. Reporting outcomes of recruiting respondents in case-control studies in a study of thyroid cancer in western Washington*

Units selected from sampling frame	Number
Random digit dialing screening phase	
Total	6741
Ineligible sampling unit	
Total	3589
Business, fax, government	1937
Nonworking numbers	1436
Institution, group quarters, dataline	216
Unable to determine eligibility	
Total	431
Unknown if residential	274
Residential, unknown if individual eligible	157
Answering machine on all attempts	56
Refusal to answer questions on eligibility	76
Other (language barrier)	25
Respondent not eligible	
Total	1983
Age	1749
County	216
Language	18
Respondent screened and eligible, total	738
In-Person interviews of eligible women	
Total	738
Unable to determine eligibility	0
Respondent not eligible	
Total	1
Prior thyroid cancer	1
Respondent screened and eligible	
Total	737
Not interviewed (refused)	163
Interviewed	574

* adapted from Olson et al. (2002)

the data. If assumptions implicit in the planned analysis methods are violated, alternative statistical methods must be considered. Appropriate and careful statistical analysis is integral to good epidemiological practice. A description of basic methods of analysis for epidemiological study designs can be found in Part II of this handbook "Statistical Methods in Epidemiology," and in most

intermediate textbooks of epidemiology (Rothman and Greenland 1998; Szklo and Nieto 2000). Some of the key issues underlying the analysis of cohort and case-control studies are summarized in Chaps. I.5 and I.6 of this handbook and in a two-volume series published by the International Agency for Cancer Research (Breslow and Day 1980, 1987). The finer points of analysis, however, are study-specific. For this reason, it is crucial that data analysis be conducted by personnel with the necessary training and experience in statistical methods.

Once data analysis is complete, ways to check the analysis include independently reproducing the tabulations and statistical calculations from the original data, and checking different tables for consistency of the denominators. All data reduction and statistical procedures should be documented to facilitate review at a later date.

The results of any study are associated with some degree of uncertainty. To the extent possible, these uncertainties should be quantified and accounted for, or, at the minimum, characterized quantitatively. In an analysis of risk factors for coronary disease in the Framingham Heart Study, estimates of risk increased for factors measured with substantial error after correction for uncertainty (e.g. serum cholesterol), whereas risk estimates tended to remain unchanged for risk factors with little or no error, such as body mass index (Rosner et al. 1992).

The analysis of study data is followed by the task of interpreting the study results. An observed association might be due to statistical artifact, due to bias or confounding, or be truly causal. The use of statistical significance alone to guide inference is not recommended (Goodman 1999a; Goodman 1999b). If one hundred truly null associations were tested at the $\alpha = 0.05$ level, five of these associations would be significant due to chance alone. Moreover, an association might be confounded by one or more variables, or could be biased due to systematic flaws in the design or conduct of the study.

Following adequate consideration of chance, confounding and bias (cf. Chap. I.9 of this handbook), the determination of whether an association is causal will also depend on *temporality*, the *strength of the association*, the presence or absence of a *dose-response relationship*, *consistency* with prior literature, and *biological plausibility* (Gordis 2000; US Department of Health Education and Welfare (DHEW) 1964).

If an exposure is believed to cause the disease in question, this exposure must occur before the disease develops. *Temporality* is easier to establish for prospective cohort studies for which exposure information prior to disease outcome is available. For cross-sectional or case-control studies, exposure information is usually collected concurrently with disease information or has to be recreated from historical records of exposure, making the assessment of temporality more difficult.

In general, the larger the magnitude of the association, the more likely it is that the relationship between the exposure and disease is causal. In epidemiologic studies, the *strength of the association* is usually measured by the relative risk or odds ratio.

If it can be demonstrated that increasing the dose of an agent is associated with increased occurrence of disease in a well-defined relationship, this provides more

evidence for causality. The absence of a *dose-response relationship*, however, does not preclude a causal relationship; since it is possible that no disease develops until a certain exposure level is reached, after which disease can occur ("threshold effect").

Consistent replication of a finding in different study populations provides further evidence for a causal relationship. However, it is possible that an association only occurs in certain population sub-groups, in which case it might be seen in some populations but not others.

Before concluding that an association is causal, it is important to consider *biological plausibility*. While it is possible that epidemiological studies can detect associations which are not yet understood on a biological level, attempting to understand how the exposure might cause the disease in question is nonetheless worthwhile.

Once the results of a study have been finalized, the investigators should consider how they plan to communicate the results, and to whom. Groups that should be informed, in general, are the study personnel, study participants, and scientific community. If the results of a study warrant immediate action, health care providers and policy makers should also be alerted. While it is important that the media is informed of the results of studies that have relevance to the general public, it is generally prudent to wait until the study is published in a peer-reviewed journal, since the process of critical review of a study allows for the identification and correction of key flaws.

A typical study report consists of the following sections: introduction, methods, results, and discussion (Table 13.6).

Regardless of the audience for the report, results should always be placed in context of the uncertainties and limitations associated with the findings. Describing results in terms of adjectives such as "definitive" or "conclusive" should be avoided. Too often, associations that receive much publicity to begin with have to be rescinded in light of further research.

Concise, simple language aids clarity of presentation. For written reports, adequately labeled tables and figures should be used to summarize information when possible. Information presented in tables should not be merely repeated in the text without additional interpretation.

It is important that results of well-designed studies are reported regardless of whether findings are negative or positive. The tendency for positive findings to be highlighted, both in terms of submission and final publication, biases the perception of the true association between exposure and outcome. This is especially problematic in the context of meta-analyses (cf. Chap. II.7 of this handbook) that attempt to quantitatively summarize published studies. A bias towards publishing positive findings results in a biased estimation of overall risk (Easterbrook et al. 1991; Egger and Smith 1998; Ioannidis 1998; Thornton and Lee 2000).

Studies with substantive findings on a research question may have implications for policies related to public health. Researchers may appropriately highlight such findings in their reports, often at the conclusion of the discussion, commenting on the extent to which new knowledge has been generated with policy implica-

Table 13.6. Guidelines for preparation of a study report*

Introduction	
Review study rationale	Describe importance of problem. Biological plausibility. How does this study add to existing literature?
State hypotheses	Specify interactions of a priori interest.
Methods	
Describe study population	Methods of recruitment. Inclusion and exclusion criteria.
Describe data collection	Include accuracy and reliability of procedures, and quality control measures.
State criteria for identification of confounders	
Describe statistical methods	Justify categorization of study variables. State assumptions of selected model.
Results	
Describe rates of participation or response	
Provide descriptive data	Frequency distributions, means, unadjusted differences. Stratify by variables of interest e.g. age, sex. Quality control measures.
Present results of model	Use most parsimonious model. Additive and multiplicative interactions, if present.
Tables and figures	Should be self-explanatory. Use informative labels, and units.
Discussion	
Review main study results	Compare and contrast with published literature.
Describe strengths and limitations of study	
Assess bias and confounding	How much would study results be affected by bias/confounding?
Address uncertainty	How precise are the study estimates, given misclassification?
Clinical, public health policy implications.	If strength and impact of study results warrants.
Future directions	How to improve on study, build on findings.

* adapted from Szklo and Nieto (2000)

tions. There has been substantial debate among epidemiologic researchers as to whether publications should also make policy recommendations (Samet 2000). In general, policy recommendations should not be made in publications providing research findings, particularly within the constraints of the policy expertise of most researchers and the space that can be devoted to such discussion in an article.

13.5.3 Storage and Retrieval of Data

Commitment to an epidemiological study does not end with the publication of the final papers. After the study is completed, sufficient material should be stored to allow future sharing of the data or auditing of the study. An index of all stored study materials should be created, along with a description of where they can be located. Materials that should be considered for archiving include source data and specimens, laboratory or research notebooks, and the study protocol. Also included should be the final study report, computer data files, copies of computer programs and statistical procedures that were used in analysis, and any printouts of analyses that formed the basis of results included in the final report (Freedland and Carney 1992; The Chemical Manufacturers Association's Epidemiology Task Force 1991; US Environmental Protection Agency (EPA) 1989). If applicable, study forms and related forms should be destroyed in accordance with local statutes and medical records. In order to ensure safety and confidentiality of study materials, storage should be in a physically secure place with limited access.

Periodic checking of stored material is recommended, to ensure that necessary updates have been made and to avoid unnecessary clutter. Original records can be transferred to microfilm for storage purposes, to conserve space. If microfilm is used, the original records should be retained until the microfilm is checked for proper identification and legibility. For very large studies, electronic storage of study data might make sense given space and cost limitations.

13.6 Conclusions

The field of epidemiology has been growing rapidly, with a vast number of epidemiologic studies published every year. A search for "Epidemiology" in the PUBMED database yielded 287 references for the year 1964. A similar search for the year 2002 yielded 46,658 references (Fig. 13.3). The results of many of these studies, however, are inconsistent. These inconsistencies are sometimes due to chance, but often can be ascribed to the variable quality of studies with respect to design, conduct, analysis or dissemination.

As a consequence of the inconsistent results reported by epidemiological studies, many consumers of epidemiological research including clinicians, policymakers, and the general public, are dismissive of new findings. The importance of

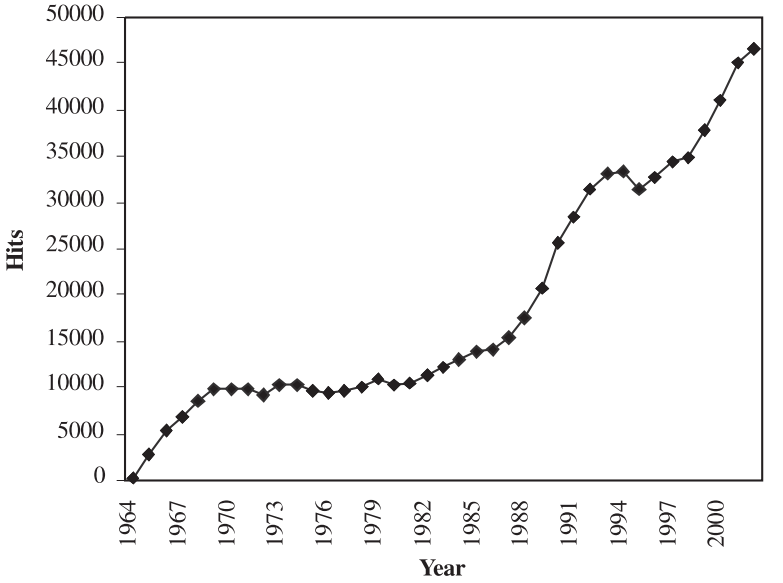

Figure 13.3. Number of references to "Epidemiology" in the PUBMED database, 1964–2002 (no delimiters)

a widespread effort to follow good epidemiologic practice and implement rigorous quality assurance and quality control procedures cannot be overstated.

Even as this chapter is being written, the methods of data collection, processing and storage are changing rapidly as technological innovations emerge. However, the basic principles of good epidemiologic practice, data quality assurance and control will not change. The increasing use of e-mail or web-based questionnaires may reduce data error due to data transfer from paper to electronic files, for example, but errors due to poor questionnaire design or data entry (to name just a few sources of error) will still exist. Similarly, electronic processing and storage of data might be helpful in identifying unusual values, but study investigators will still need to review, interpret, and correct these errors.

In this chapter, we have reviewed quality assurance and quality control activities pertinent to the planning, conduct and reporting of a study. The mental exercise of "optimizing" the dials on the datascope can be useful while conducting epidemiological studies, and when considering the results of already published studies. As high quality research becomes the norm, the field of epidemiology will gain more respect among fellow scientists, policy-makers, and the public.

References

Agresti A (1990) Categorical data analysis. John Wiley and Sons, Hoboken, NJ

Altman D, Bland J (1983) Measurements in medicine: the analysis of method comparison studies. Statistician 32:307–317

American Association for Public Opinion Research. Standard definitions (2000) Final dispositions of case codes and outcome rates for surveys. American Association for Public Opinion Research, Ann Arbor, MI

Armstrong B, White E (1992) Principles of exposure measurement in epidemiology. Oxford University Press, New York

Arts DG, De Keizer NF, Scheffer GJ (2002) Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc 9:600–611

Asch DA, Jedrziewski MK, Christakis NA (1997) Response rates to mail surveys published in medical journals. J Clin Epidemiol 50:1129–1136

Asch DA, Christakis NA, Ubel PA (1998) Conducting physician mail surveys on a limited budget. A randomized trial comparing $2 bill versus $5 bill incentives. Med Care 36:95–99

Ascherio A, Stampfer MJ, Colditz GA, Rimm EB, Litin L, Willet WC (1992) Correlations of vitamin A and E intakes with the plasma concentrations of carotenoids and tocopherols among American men and women. J Nutr 122: 1792–1801

Baer A, Saroiu S, Koutsky LA (2002) Obtaining sensitive data through the Web: an example of design and methods. Epidemiol 13:640–645

Barnett V, Lewis T (1994) Outliers in statistical data. John Wiley and Sons, Hoboken, N.J.

Berkanovic E (1980) The effect of inadequate language translation on Hispanics' responses to health surveys. Am J Public Health 80:1273–1276

Bisgard KM, Folsom AR, Hong CP, Sellers TA (1994) Mortality and cancer rates in nonrespondents to a prospective study of older women: 5-year follow-up. Am J Epidemiol 139:990–1000

Blackmore CC, Richardson ML, Linnau KF, Schwed AM, Lomoschitz FM, Escobedo EM, Hunter JC, Jurkovich GJ, Cummings P (2003) Web-based image review and data acquisition for multiinstitutional research. AJR Am J Roentgenol 180:1243–1246

Brenner H, Gefeller O (1997) Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med 16:981–991

Breslow N, Day N (1980) Statistical methods in cancer research. Volume I – The analysis of case-control studies. International Agency for Research on Cancer, Lyon

Breslow N, Day N (1987) Statistical methods in cancer research. Volume II – The design and analysis of cohort studies. International Agency for Research on Cancer, Lyon

Bryant AH, Reinert A (2001) Epidemiology in the legal arena and the search for truth. Am J Epidemiol 154:S27–S35

Canner PL, Krol WF, Forman SA (1983) The Coronary Drug Project. External quality control programs. Control Clin Trials 4:441–466

Canner PL, Borhani NO, Oberman A, Cutler J, Prineas RJ, Langford H, Hooper FJ (1991) The Hypertension Prevention Trial: assessment of the quality of blood pressure measurements. Am J Epidemiol 134:379–392

Centers for Disease Control and Prevention (CDC) (1999) BRFSS summary quality control report. Centers for Disease Control and Prevention, Atlanta, GA

Cherrie J, Schneider T (1998) Validation of a new method for structured subjective assessment of past concentrations. Annals Occup Hyg 43:235–245

Cherrie J, Krantz S, Schneider T, Ohberg I, Kamstrup O, Linander W (1987) An experimental simulation of an early rock wool/slag wool production process. Ann Occup Hyg 31:583–593

Choi BC, Pak AW, Purdham JT (1990) Effects of mailing strategies on response rate, response time, and cost in a questionnaire study among nurses. Epidemiol 1:72–74

Christiansen DH, Hosking JD, Dannenberg AL, Williams OD (1990) Computer-assisted data collection in multicenter epidemiologic research. The Atherosclerosis Risk in Communities Study. Control Clin Trials 11:101–115

Cicchetti DV, Feinstein AR (1990) High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 43:551–558

Clement DL, De Buyzere ML, De Bacquer DA, de Leeuw PW, Duprez DA, Fagard RH, Gheeraert PJ, Missault LH, Braun JJ, Six RO, Van Der NP, O'Brien E (2003) Prognostic value of ambulatory blood-pressure recordings in patients with treated hypertension. N Engl J Med 348:2407–2415

Clive RE, Ocwieja KM, Kamell L, Hoyler SS, Seiffert JE, Young JL, Henson DE, Winchester DP, Osteen RT, Menck HR (1995) A national quality improvement effort: cancer registry data. J Surg Oncol 58:155–161

Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement of partial credit. Psychological Bulletin 70:213–220

Collins RL, Ellickson PL, Hays RD, McCaffrey DF (2000) Effects of incentive size and timing on response rates to a follow-up wave of a longitudinal mailed survey. Eval Rev 24:347–363

Comstock GW, Tockman MS, Helsing KJ, Hennesy KM (1979) Standardized respiratory questionnaires: comparison of the old with the new. Am Rev Respir Dis 119:45–53

Cook RR (1991) Overview of good epidemiologic practices. J Occup Med 33: 1216–1220

Cooper GR (1986) The importance of quality control in the Multiple Risk Factor Intervention Trial. Control Clin Trials 7:3pp

Cottler LB, Zipp JF, Robins LN, Spitznagel EL (1987) Difficult-to-recruit respondents and their effect on prevalence estimates in an epidemiologic survey. Am J Epidemiol 125:329–339

Crombie IK, Irving JM (1986) An investigation of data entry methods with a personal computer. Comput Biomed Res 19:543–550

Dawber TR, Meadors GF, Moore FE, Jr. (1951) Epidemiological approaches to heart disease: The Framingham Study. Am J Public Health 41:279–286

Day S, Fayers P, Harvey D (1998) Double data entry: what value, what price? Control Clin Trials 19:15–24

Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 12:142S–158S

Dillman D (1978) Mail and telephone surveys: the total design method. John Wiley and Sons, New York

Dischinger P, DuChene AG (1986) Quality control aspects of blood pressure measurements in the Multiple Risk Factor Intervention Trial. Control Clin Trials 7:137S–157S

Doody MM, Sigurdson AS, Kampa D, Chimes K, Alexander BH, Ron E, Tarone RE, Linet MS (2003) Randomized trial of financial incentives and delivery methods for improving response to a mailed questionnaire. Am J Epidemiol 157:643–651

Dosemeci M, Rothman N, Yin SN, Li GL, Linet M, Wacholder S, Chow WH, Hayes RB (1997) Validation of benzene exposure assessment. Ann N Y Acad Sci 837:114–121

DuChene AG, Hultgren DH, Neaton JD, Grambsch PV, Broste SK, Aus BM, Rasmussen WL (1986) Forms control and error detection procedures used at the Coordinating Center of the Multiple Risk Factor Intervention Trial (MRFIT). Control Clin Trials 7:34S–45S

Eaker S, Bergstrom R, Bergstrom A, Adami HO, Nyren O (1998) Response rate to mailed epidemiologic questionnaires: a population-based randomized trial of variations in design and mailing routines. Am J Epidemiol 147:74–82

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR (1991) Publication bias in clinical research. Lancet 337:867–872

Edwards P, Roberts I, Clarke M, DiGuiseppi C, Pratap S, Wentz R, Kwan I (2002) Increasing response rates to postal questionnaires: systematic review. Br Med J 324:1183

Edwards S, Slattery ML, Mori M, Berry TD, Caan BJ, Palmer P, Potter JD (1994) Objective system for interviewer performance evaluation for use in epidemiologic studies. Am J Epidemiol 140:1020–1028

Egger M, Smith GD (1998) Bias in location and selection of studies. Br Med J 316:61–66

Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 43:543–549

Fleiss JL (1981) Statistical methods for rates and proportions, 2nd edn. John Wiley and Sons, New York

Fleming TR (1993) Data monitoring committees and capturing relevant information of high quality. Stat Med 12:565–570

Fowler F, Mangione T (1986) Reducing interviewer effects on health survey data. Center for Survey Research, University of Massachusetts, Boston, MA

Fowler F, Mangione T (1990) Standardized survey interviewing: minimizing interviewer-related error. Sage Publications, Newberry Park, CA

Freedland KE, Carney RM (1992) Data management and accountability in behavioral and biomedical research. Am Psychol 47:640–645

Gassman JJ, Owen WW, Kuntz TE, Martin JP, Amoroso WP (1995) Data quality assurance, monitoring, and reporting. Control Clin Trials 16:104S–136S

Gibson PJ, Koepsell TD, Diehr P, Hale C (1999) Increasing response rates for mailed surveys of Medicaid clients and other low-income populations. Am J Epidemiol 149:1057–1062

Gilbart E, Kreiger N (1998) Improvement in cumulative response rates following implementation of a financial incentive. Am J Epidemiol 148:97–99

Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J (1996) Chart reviews in emergency medicine research: Where are the methods? Ann Emerg Med 27:305–308

Gissler M, Teperi J, Hemminki E, Merilainen J (1995) Data quality after restructuring a national medical registry. Scand J Soc Med 23:75–80

Goldberg J, Gelfand HM, Levy PS (1980) Registry evaluation methods: a review and case study. Epidemiol Rev 2:210–220

Goldman LR (2001) Epidemiology in the regulatory arena. Am J Epidemiol 154: S18–S26

Goodman SN (1999a) Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med 130:1005–1013

Goodman SN (1999b) Toward evidence-based medical statistics: 1. The *P* value fallacy. Ann Intern Med 130:995–1004

Gordis L (2000) Epidemiology, 2nd edn. W.B. Saunders, Philadelphia

Greenbaum DS, Bachmann JD, Krewski D, Samet JM, White R, Wyzga RE (2001) Particulate air pollution standards and morbidity and mortality: case study. Am J Epidemiol 154:S78–S90

Halpern SD, Ubel PA, Berlin JA, Asch DA (2002) Randomized trial of 5 dollars versus 10 dollars monetary incentives, envelope size, and candy to increase physician response rates to mailed questionnaires. Med Care 40:834–839

Hawkins N, Evans J (1989) Subjective estimation of toluene exposures: a calibration study of industrial hygienists. Appl Ind Hygiene 4:61–68

Hearst N, Hulley SB (1988) Using secondary data. In: Hulley SB, Cummings SR (eds) Designing clinical research. Williams & Wilkins, Baltimore, MD, pp. 53–62

Heilbrun LK, Nomura A, Stemmermann GN (1991) The effects of non-response in a prospective study of cancer: 15-year follow-up. Int J Epidemiol 20:328–338

Hill J (2003) Certification in the Cardiovascular Health Study. Personal communication with Rajaraman P

Hilner JE, McDonald A, Van Horn L, Bragg C, Caan B, Slattery ML, Birch R, Smoak CG, Wittes J (1992) Quality control of dietary data collection in the CARDIA study. Control Clin Trials 13:156–169

Hoffman SC, Burke AE, Helzlsouer KJ, Comstock GW (1998) Controlled trial of the effect of length, incentives, and follow-up techniques on response to a mailed questionnaire. Am J Epidemiol 148:1007–1011

Holford TR, Stack C (1995) Study design for epidemiologic studies with measurement error. Stat Methods Med Res 4:339–358

Horbar JD, Leahy KA (1995) An assessment of data quality in the Vermont-Oxford Trials Network database. Control Clin Trials 16:51–61

Hosking JD, Rochon J (1982) A comparison of techniques for detecting and preventing key-field errors. Proceedings of the Statistical Computing Section. 82–87. American Statistical Association, Washington, D.C.

Hosking JD, Newhouse MM, Bagniewska A, Hawkins BS (1995) Data collection and transcription. Control Clin Trials 16:66S–103S

Hunt JR, White E (1998) Retaining and tracking cohort study members. Epidemiol Rev 20:57–70

International Organization for Standardization (2003) ISO 9000:2000, ISO Technical Committee ISO/TC 176

Ioannidis JP (1998) Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 279:281–286

James J, Bolstein R (1992) Large monetary incentives and their effect on mail survey response rates. Public Opinion Quarterly 56:442–453

John EM, Savitz DA (1994) Effect of a monetary incentive on response to a mail survey. Ann Epidemiol 4:231–235

Johnstone FD, Brown MC, Campbell D, MacGillivray I (1981) Measurement of variables: data quality control. Am J Clin Nutr 34:804–806

Kaaks R, Ferrari P, Ciampi A, Plummer M, Riboli E (2002) Uses and limitations of statistical accounting for random error correlations, in the validation of dietary questionnaire assessments. Public Health Nutr 5:969–976

Kalantar JS, Talley NJ (1999) The effects of lottery incentive and length of questionnaire on health survey response rates: a randomized study. J Clin Epidemiol 52:1117–1122

Kannel WB (2000) Risk stratification in hypertension: new insights from the Framingham Study. Am J Hypertens 13:3S–10S

Kellerman SE, Herold J (2001) Physician response to surveys. A review of the literature. Am J Prev Med 20:61–67

Kessler RC, Little RJ, Groves RM (1995) Advances in strategies for minimizing and adjusting for survey nonresponse. Epidemiol Rev 17:192–204

Kiesler S, Sproull L (1986) Response effects in the electronic survey. Public Opinion Quarterly 50:402–413

Kipen HM, Cody RP, Goldstein BD (1989) Use of longitudinal analysis of peripheral blood counts to validate historical reconstructions of benzene exposure. Environ Health Perspect 82:199–206

Kjelsberg MO, Cutler JA, Dolecek TA (1997) Brief description of the Multiple Risk Factor Intervention Trial. Am J Clin Nutr 65:191S–195S

Knatterud GL, Rockhold FW, George SL, Barton FB, Davis CE, Fairweather WR, Honohan T, Mowery R, O'Neill R (1998) Guidelines for quality assurance in multicenter trials: a position paper. Control Clin Trials 19:477–493

Krewski D, Burnett RT, Goldberg MS, Hoover K, Siemiatycki J, Abrahamowicz M, White WH (2000) Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of particulate air pollution and mortality. Investigators' reports parts I and II. Health Effects Institute, Cambridge, MA

Kromhout H, Oostendorp Y, Heederik D, Boleij JS (1987) Agreement between qualitative exposure estimates and quantitative exposure measurements. Am J Ind Med 12:551–562

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Ling PM, Glantz SA (2002) Using tobacco-industry marketing research to design more effective tobacco-control campaigns. JAMA 287:2983–2989

Little RE, Davis AK (1984) Effectiveness of various methods of contact and reimbursement on response rates of pregnant women to a mail questionnaire. Am J Epidemiol 120:161–163

Maclure M, Schneeweiss S (2001) Causation of bias: the episcope. Epidemiol 12: 114–122

Maclure M, Willett WC (1987) Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 126:161–169

Maheux B, Legault C, Lambert J (1989) Increasing response rates in physicians' mail surveys: an experimental study. Am J Public Health 79:638–639

Martinson BC, Lazovich D, Lando HA, Perry CL, McGovern PG, Boyle RG (2000) Effectiveness of monetary incentives for recruiting adolescents to an intervention trial to reduce smoking. Prev Med 31:706–713

Maudsley G, Williams EM (1999) What lessons can be learned for cancer registration quality assurance from data users? Skin cancer as an example. Int J Epidemiol 28:809–815

McQuade CE, Kutvirt DM, Brylinski DA, Samet JM (1983) A tracking system for conducting epidemiological case-control studies. Comput Programs Biomed 16:149–153

Meinert CL, Tonascia S (1986) Controlled clinical trials: design, conduct, and analysis. Oxford University Press, New York

Moorman PG, Newman B, Millikan RC, Tse CK, Sandler DP (1999) Participation rates in a case-control study: the impact of age, race, and race of interviewer. Ann Epidemiol 9:188–195

Mullooly JP (1990) The effects of data entry error: an analysis of partial verification. Comput Biomed Res 23:259–267

Neaton JD, DuChene AG, Svendsen KH, Wentworth D (1990) An examination of the efficiency of some quality assurance methods commonly employed in clinical trials. Stat Med 9:115–123

Olson SH (2001) Reported participation in case-control studies: changes over time. Am J Epidemiol 154:574–581

Olson SH, Voigt LF, Begg CB, Weiss NS (2002) Reporting participation in case-control studies. Epidemiol 13:123–126

Paolo AM, Bonaminio GA, Gibson C, Partridge T, Kallail K (2000) Response rate comparisons of e-mail- and mail-distributed student evaluations. Teach Learn Med 12:81–84

Parkes R, Kreiger N, James B, Johnson KC (2000) Effects on subject response of information brochures and small cash incentives in a mail-based case-control study. Ann Epidemiol 10:117–124

Perneger TV, Etter JF, Rougemont A (1993) Randomized trial of use of a monetary incentive and a reminder card to increase the response rate to a mailed health survey. Am J Epidemiol 138:714–722

Post W, Kromhout H (1991) Semiquantitative estimates of exposure to methylene chloride adn styrene: the influence of quantitative exposure data. Applied Occupational and Environmental Hygiene 6:197–204

Prud'homme GJ, Canner PL, Cutler JA (1989) Quality assurance and monitoring in the Hypertension Prevention Trial. Hypertension Prevention Trial Research Group. Control Clin Trials 10:84S–94S

Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, Wahl PW (1997) The Sleep Heart Health Study: design, rationale, and methods. Sleep 20:1077–1085

Reisch LM, Fosse JS, Beverly K, Yu O, Barlow WE, Harris EL, Rolnick S, Barton MB, Geiger AM, Herrinton LJ, Greene SM, Fletcher SW, Elmore JG (2003) Training, quality assurance, and assessment of medical record abstraction in a multisite study. Am J Epidemiol 157:546–551

Rhodes SD, Bowie DA, Hergenrather KC (2003) Collecting behavioural data using the world wide web: considerations for researchers. J Epidemiol Community Health 57:68–73

Rosner B, Spiegelman D, Willett WC (1992) Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. Am J Epidemiol 136:1400–1413

Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott-Raven, Philadelphia

Sacks FM, Handysides GH, Marais GE, Rosner B, Kass EH (1986) Effects of a low-fat diet on plasma lipoprotein levels. Arch Intern Med 146:1573–1577

Samet JM (2000) Epidemiology and policy: the pump handle meets the new millennium. Epidemiol Rev 22:145–154

Samet JM, Lee NL (2001) Bridging the gap: perspectives on translating epidemiologic evidence into policy. Am J Epidemiol 154:S1–S3

Samet JM, Zeger SL, Kelsall JE, Xu J, Kalkstein LS (1997) Particulate air pollution and daily mortality: analyses of the effects of weather and multiple air pollutants (The Phase IB Report of the Particle Epidemiology Evaluation Project). Health Effects Institute, Cambridge, MA

Schweitzer M, Asch DA (1995) Timing payments to subjects of mail surveys: cost-effectiveness and bias. J Clin Epidemiol 48:1325–1329

Shahar E, Folsom AR, Jackson R (1996) The effect of nonresponse on prevalence estimates for a referent population: insights from a population-based cohort study. Atherosclerosis Risk in Communities (ARIC) Study Investigators. Ann Epidemiol 6:498–506

Shaw MJ, Beebe TJ, Jensen HL, Adlis SA (2001) The use of monetary incentives in a community survey: impact on response rates, data quality, and cost. Health Serv Res 35:1339–1346

Silver RC, Holman EA, McIntosh DN, Poulin M, Gil-Rivas V (2002) Nationwide longitudinal study of psychological responses to September 11. JAMA 288: 1235–1244

Slattery ML, Edwards SL, Caan BJ, Kerber RA, Potter JD (1995) Response rates among control subjects in case-control studies. Ann Epidemiol 5:245–249

Sorensen HT, Sabroe S, Olsen J (1996) A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol 25:435–442

Spiegelman D, Schneeweiss S, McDermott A (1997) Measurement error correction for logistic regression models with an "alloyed gold standard". Am J Epidemiol 145:184–196

Spry VM, Hovell MF, Sallis JG, Hofsteter CR, Elder JP, Molgaard CA (1989) Recruiting survey respondents to mailed surveys: controlled trials of incentives and prompts. Am J Epidemiol 130:166–172

Steeh C (1981) Trends in nonresponse rates 1952–1979. Public Opinion Quarterly 45:40–57

Stram DO, Langholz B, Huberman M, Thomas DC (1999) Correcting for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado Plateau uranium miners cohort. Health Phys 77:265–275

Szklo M, Nieto FJ (2000) Epidemiology: beyond the basics. Aspen, Gaithersburg, MD

The Chemical Manufacturers Association's Epidemiology Task Force (1991) Guidelines for good epidemiological practices for occupational and environmental epidemiologic research. J Occup Med 33:1221–1229

Thompson WD (1990) Kappa and attenuation of the odds ratio. Epidemiol 1: 357–369

Thompson WD, Walter SD (1988) A reaapraisal of the kappa coefficient. J Clin Epidemiol 41:949–958

Thornton A, Lee P (2000) Publication bias in meta-analysis: its causes and consequences. J Clin Epidemiol 53:207–216

Tielemans E, Heederik D, Burdorf A, Vermeulen R, Veulemans H, Kromhout H, Hartog K (1999) Assessment of occupational exposures in a general population: comparison of different methods. Occup Environ Med 56:145–151

Turpin J, Rose R, Larsen B (2003) An adaptable, transportable web-based data acquisition platform for clinical and survey-based research. J Am Osteopath Assoc 103:182–186

US Department of Health and Human Services (2001) Application of a Public Health Service Grant. PHS 398. Public Health Service

US Department of Health Education and Welfare (DHEW) (1964) Smoking and health. Report of the Advisory Committee to the Surgeon General. DHEW Publication No. [PHS] 1103. U.S. Government Printing Office, Washington, DC

US Department of Health Education and Welfare (DHEW) (1973) Final report of the Tuskegee Syphilis STudy Ad Hoc Advisory Panel. US Public Health Service, Washington, D.C.

US Environmental Protection Agency (EPA) (1989) Toxic substances control act (TSCA): good laboratory practice standards. 40 CFR Part 792, 34034–34050

Vantongelen K, Rotmensz N, van der Schueren E (1989) Quality control of validity of data collected in clinical trials. EORTC Study Group on Data Management (SGDM). Eur J Cancer Clin Oncol 25:1241–1247

Vardeman SB, Jobe JM (1999) Statistical quality assurance methods for engineers. John Wiley and Sons, Hoboken, N.J.

Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992) Selection of controls in case-control studies. I. Principles. Am J Epidemiol 135:1019–1028

Wacholder S, Armstrong B, Hartge P (1993) Validation studies using an alloyed gold standard. Am J Epidemiol 137:1251–1258

Wallace JM, Jr., Bachman JG, O'Malley PM, Johnston LD, Schulenberg JE, Cooper SM (2002) Tobacco, alcohol, and illicit drug use: racial and ethnic differences among U.S. high school seniors, 1976–2000. Public Health Rep 117 Suppl 1:S67–S75

White E, Hunt JR, Casso D (1998) Exposure measurement in cohort studies: the challenges of prospective data collection. Epidemiol Rev 20:43–56

Whitney CW, Lind BK, Wahl PW (1998) Quality assurance and quality control in longitudinal studies. Epidemiol Rev 20:71–80

Willett WC, Stampfer MJ, Underwood BA, Speizer FE, Rosner B, Hennekens CH (1983) Validation of a dietary questionnaire with plasma carotenoid and alpha-tocopherol levels. Am J Clin Nutr 38:631–639

Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, Hennekens CH, Speizer FE (1985) Reproducibiltiy and validity of a semiquantitative food frequency questionaire. Am J Epidemiol 122:51–65

Wright P, Haybittle J (1979a) Design of forms for clinical trials (1). Br Med J 2: 529–530

Wright P, Haybittle J (1979b) Design of forms for clinical trials (2). Br Med J 2:590–592

Wright P, Haybittle J (1979c) Design of forms for clinical trials (3). Br Med J 2: 650–651

Wyatt J (1995) Acquisition and use of clinical data for audit and research. J Eval Clin Pract 1:15–27