Exposure Measurement in Cohort Studies: The Challenges of Prospective Data Collection

Emily White,1, 2 Julie R. Hunt,1 and Deborah Casso1, 2

INTRODUCTION

Collection of accurate exposure information on study participants is a major objective of all epidemiologic studies. The term exposure is used here to include not only exposures to exogenous agents that may be causally related to disease, but also socioeconomic factors, health habits, endogenous factors, and factors that may confound or modify the primary exposure-disease relation. Because many of the issues in exposure measurement are similar for case-control and cohort studies (1-3), this presentation attempts to focus on the unique challenges in measurement that arise from certain characteristics of many (but not all) cohort studies; these include the prospective, repeated exposure assessment and the larger sample size, which lead to greater need for cost considerations in data collection. We review methods of exposure measurement in cohort studies and issues of data collection costs, including selection of cost-efficient measurement methods and sampling approaches to reduce costs. Issues relating to prospective data collection are covered, including choice of a reference period, frequency of exposure updates, changes in the instrument over time, and use of repeated measures. The final focus is on sources of exposure measurement error, and approaches to reduce such measurement error through the design, execution, and analysis of the study. These are key issues, because exposure measurement error can lead to substantial bias in the estimated relative risk for the exposure-disease relation (1, 4, 5).

MEASUREMENT METHODS IN COHORT STUDIES

The first measurement issue in designing a cohort study is usually the selection of the data collection method(s). While selecting an accurate method should be the primary concern, cost considerations often limit what is practicable.

Selection of a measurement method

A large range of exposure measurement methods have been used in prospective cohort studies (examples are presented in table 1). Prospective cohort studies have used mailed self-administered questionnaires, interviewer-administered questionnaires, measures of blood or other tissues, physical measures, medical tests (such as electrocardiography), use of medical or other exposure records, and/or measures of the environment (e.g., air or water sampling) (table 1). Many studies use multiple methods. For example, the Cardiovascular Health Study uses home interviews and clinic visits, which include interviewer- and self-administered questionnaires, physical tests (for anthropometrics), blood measures, medical tests (blood pressure, electrocardiography, ultrasonography), and abstraction of prescription labels of medication use (27). In the study of Colorado Plateau uranium miners, Wagoner et al. (9) used physical examinations and detailed occupational questionnaires, as well as longitudinal measures of environmental radiation from each mine. Storage of blood or other sources of DNA for future studies of biomarkers has become an important component of several large cohort studies, including the Women's Health Initiative Observational Study (28) and the Nurses' Health Study (15, 16). Diaries (e.g., food records) have rarely been used in cohort studies because of the costs of training subjects and of coding the diary information. However, diaries could be more cost effective if the coding were limited to a subsample of participants (see "Measurement of exposures on a sample of the full cohort to reduce costs" below).

A major consideration in the selection of a method is the validity of the method. Although issues surrounding validity are often specific for each type of exposure, one general issue is whether a current measure (or prospectively repeated measures of current exposure) is appropriate or whether a measure of past...
### TABLE 1. Examples of exposure assessment methods in selected prospective cohort studies

<table>
<thead>
<tr>
<th>Study (reference(s))</th>
<th>Year begun</th>
<th>Main focus</th>
<th>Study population</th>
<th>Size</th>
<th>Baseline exposure instruments*</th>
<th>Follow-up exposure instruments*</th>
<th>Frequency exposure follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Framingham Heart Study (6-8)</td>
<td>1948</td>
<td>Risk factors for cardiovascular disease</td>
<td>Residents of Framingham, MA, ages 28–62 years</td>
<td>5,209</td>
<td>Interview, clinic examination (PE, lab, tests)</td>
<td>Interview, clinic examination (PE, lab, tests)</td>
<td>Every 2 years</td>
</tr>
<tr>
<td>Colorado Plateau Uranium Miners Study (9)</td>
<td>1950</td>
<td>Occupational risk factors for cancer</td>
<td>White, male underground uranium miners, Colorado Plateau</td>
<td>3,415</td>
<td>Clinic examination (PE), questionnaire, environmental measures (airborne radiation)</td>
<td>Questionnaire (in-person or mailed), environmental measures (airborne radiation)</td>
<td>Triennial</td>
</tr>
<tr>
<td>American Cancer Society: Cancer Prevention Study I (10)</td>
<td>1959</td>
<td>Cigarette smoking and cancer mortality</td>
<td>US men and women aged 30 years and older</td>
<td>1,045,087</td>
<td>Self-administered questionnaire delivered by volunteers</td>
<td>Self-administered questionnaire</td>
<td>Every 2 years</td>
</tr>
<tr>
<td>The Alameda County Study (11)</td>
<td>1965</td>
<td>Factors associated with health and mortality</td>
<td>Residents of Alameda County, CA, ages 16–94 years</td>
<td>6,928</td>
<td>Mailed questionnaire, telephone interview or home interview of nonrespondents</td>
<td>Mailed questionnaire, telephone interview or home interview of nonrespondents</td>
<td>At years 9, 18, 26</td>
</tr>
<tr>
<td>Honolulu Heart Program (12)</td>
<td>1965</td>
<td>Coronary heart disease and stroke in men of Japanese ancestry</td>
<td>Men of Japanese ancestry living on Oahu, HI, ages 45–65 years</td>
<td>8,006</td>
<td>Mailed questionnaire, interview, clinic examination (PE, lab, tests)</td>
<td>Same</td>
<td>Every 6 months</td>
</tr>
<tr>
<td>The Oral Contraception Study of the Royal College of General Practitioners (13, 14)</td>
<td>1968</td>
<td>Oral contraceptive use and cancer</td>
<td>Married premenopausal British women</td>
<td>47,000</td>
<td>Medical form completed by physician based on patient interview or medical record</td>
<td>Same</td>
<td>Every 6 months</td>
</tr>
<tr>
<td>Nurses' Health Study (15, 16)</td>
<td>1976</td>
<td>Originally oral contraceptive use and cancer, expanded to women's health</td>
<td>Married female US registered nurses ages 30–55 years</td>
<td>121,700</td>
<td>Mailed questionnaire</td>
<td>Mailed questionnaire, Lab (toenail sample)</td>
<td>Every 2 years</td>
</tr>
<tr>
<td>Port Pirie Cohort Study (17, 18)</td>
<td>1979</td>
<td>Lead exposure and child development</td>
<td>Infants born in Port Pirie, South Australia, 1979–1982</td>
<td>723</td>
<td>Lab (blood samples from pregnant mother and umbilical cord at birth)</td>
<td>Lab (blood samples)</td>
<td>At 6, 15, and 24 annually up and at 11–12 years</td>
</tr>
<tr>
<td>Multicenter AIDS Cohort Study (19–20)</td>
<td>1984</td>
<td>Risk factors for HIV† in gay men</td>
<td>US homosexual men ages 18–70 years</td>
<td>4,954</td>
<td>Clinic examination (PE, lab), self-administered questionnaire, interview</td>
<td>Clinic examination (PE, lab), self-administered questionnaire, interview</td>
<td>Every 6 months</td>
</tr>
<tr>
<td>Coronary Artery Risk Development in Young Adults (CARDIA) (21, 22)</td>
<td>1985</td>
<td>Risk factors for coronary heart disease in young adults</td>
<td>Black and white US men and women ages 18–30 years</td>
<td>5,115</td>
<td>Telephone interview, clinic examination (PE, lab, tests), self-administered questionnaire, interview</td>
<td>Clinic examination (PE, lab, tests), self-administered questionnaire, interview</td>
<td>At years 2, 5, 10, 15, 20, 25 and 30</td>
</tr>
<tr>
<td>New York University Women's Health Study (23)</td>
<td>1985</td>
<td>Endogenous hormones and risk of breast cancer</td>
<td>New York women ages 35–65 years</td>
<td>14,291</td>
<td>Lab, self-administered questionnaire</td>
<td>Lab</td>
<td>Annual</td>
</tr>
<tr>
<td>Iowa Women's Study (24)</td>
<td>1986</td>
<td>Cancer in women</td>
<td>Iowa women ages 55–69 years</td>
<td>41,637</td>
<td>Mailed questionnaire with self body girth measures</td>
<td>None</td>
<td></td>
</tr>
</tbody>
</table>
Exposure Measurement

45

If exposure is needed (see the section on Developing an exposure measure—time considerations, below). Blood measures, environmental measures, and physical measures generally measure only current exposure. However, if the characteristic is fixed (e.g., inherited genetic traits) or reasonably stable over long periods of time, then a current measure can be a good measure of past exposure.

Cost considerations

For large cohort studies, cost is a major concern. The two common approaches to low-cost cohort studies are record-linkage studies, in which records of exposure are linked to disease records, and studies which recruit a cohort and collect exposures by mailed questionnaires.

Numerous studies have been published in which medical, pharmacy, educational, occupational, or other records of exposure have been used (29–31), see the article in this issue by Howe (32) for other examples. Obtaining records is usually inexpensive even compared with mailed questionnaires, particularly if the cohort is selected to facilitate the availability of records (such as members of a health maintenance organization) or if the records are computerized. Moreover, use of records often allows the conduct of a cohort study in which the exposure data collection has occurred in the past (retrospective cohort study) (2). Because the exposures are recorded for each subject before disease onset, retrospective cohort studies generally have the advantages of prospective cohort studies, including less susceptibility to differential and nondifferential measurement error and to selection bias. However, records are often not available for the exposure of interest or are not sufficiently accurate. Even when records are available with reasonably accurate measures of the primary exposure, information on potential confounders is often poorly measured or is missing (1).

Mailed questionnaires are generally the least expensive of the other available methods (33, 34), and have been used as the primary exposure method in several large cohort studies, including the Nurses' Health Study and the Iowa Women’s Study (15, 16, 24). Machine reading of questionnaires using "mark-sense" or fax (digitization) technology further reduces costs. However, mailed questionnaires generally must be shorter and less complex than interviewer-administered questionnaires. For example, complex skip patterns must be avoided on mailed questionnaires.

Other innovative methods have been used to reduce the costs of large cohort studies. In some studies, the subjects themselves have served as data collectors for
physical measures or biologic specimens, tasks that usually are conducted by study staff. In the Nurses' Health Study, the subjects were mailed materials and instructions to provide a blood sample, and were also asked for toenail samples (to assess selenium exposure) (15). In the Oral Contraceptive Study of the Royal College of General Practitioners, data collection costs were reduced by soliciting physicians to recruit subjects and be the primary informants on subjects' exposures (oral contraceptive use) (13, 14). In the American Cancer Society Cohort Studies, volunteers recruited and delivered the study questionnaire to participants (10).

**Measurement of exposures on a sample of the full cohort to reduce costs**

Another approach to reduce data collection costs in cohort studies is to collect or have available the research material for all cohort members, but to measure the exposures on less than the full cohort (2, 35-38). For example, medical records might be available for the full cohort but only abstracted on a sample of cohort members, or stored blood might be available for the full cohort but only analyzed for a subsample. Several types of sampling designs can be used, each of which has advantages and limitations (2, 35-38). In a **nested case-control design**, cases are compared with a sample of controls free of disease at the time of the case's diagnosis (39, 40). In a **two-phase design**, a proportion of each of the four disease-by-exposure groups are sampled (41, 42). In a **case-cohort design**, a random subsample of the cohort is selected at baseline, and is compared with those who develop the disease of interest (43). This latter design allows the cohort subsample to be compared with more than one disease group. These designs can lead to reduced exposure measurement costs with a relatively small loss in study power.

**DEVELOPING AN EXPOSURE MEASURE—TIME CONSIDERATIONS**

**Dose and time**

After the exposure measurement method has been selected (e.g., self-administered questionnaire for physical activity), the actual instrument is selected from among those available, or a new instrument is developed. Most of the considerations in selecting or developing an instrument are similar for cohort and case-control studies. Briefly, these begin with conceptualization of the actual "active agent" that is hypothesized to be causally related to disease, identification of those exposures which do (and do not) include the agent, and conceptualization of the best construct(s) of "dose" (e.g., cumulative dose of the active agent, cumulative dose above some threshold of intensity, etc.) (1). Then the researcher identifies the items needed to capture the duration, frequency, and intensity of exposure over a specified time period, and develops the algorithm for the analytic exposure variable(s) which combines these items to yield the exposure dose construct that is expected to be causally related to disease. Conversion tables are often used in the algorithm to quantify the intensity of exposure into units of the active agent (e.g., to convert each type of physical exercise into kilocalorie expenditure). The final step is the actual phrasing of the questions for each item. Most of these issues have been covered in the context of measurement of specific exposures, e.g., diet (44), physical activity (45), drugs (46), occupational exposures (47), laboratory-based measurements (48, 49), and in books and articles on questionnaire design (50-53).

Because cohort studies are often prospective, issues relating to the time period of importance in the exposure-disease relation deserve particular consideration. Several authors discuss the importance and methods of considering the time period of exposure in addition to the dose of exposure in evaluating the effect of exposure on disease occurrence (54-58). One of the primary ways in which the dose-time effect of exposure on disease risk can be assessed is by computing cumulative dose (or average dose) only over the time period thought to be of etiologic importance (54) (see Thomas (56) for other analytic approaches). Considering the etiologically important time period can help answer questions that arise in the design of cohort studies, such as:

- Should current or past exposures be measured?
- How often should the measure of exposure be repeated prospectively (e.g., repeat questions on exercise or repeat blood draws)?
- How should exposures that change over time be handled (e.g., how does one classify a subject who was on hormone replacement therapy at baseline but not at the 2-year follow-up)?

Additionally, one must consider how to handle changes in the instrument over time. Changes might be necessary if there are new sources of the exposure (e.g., new forms of hormone replacement therapy) or if the instrument can be improved.

**Etiologically relevant time window**

Rothman (54) provides a structure with which to view the relation between timing of exposure and the occurrence of disease. Specifically, if there is a series of component causes which must occur in a sequence
to lead to disease, then there would be a limited time period during which a given exposure is causally related to disease, i.e., an *etiologically relevant time window* (top line of figure 1). After the action of that exposure is complete, then an induction period ensues during which any remaining component causes in the etiologic sequence occur. The length of the induction period therefore varies for each exposure related to the disease, with those acting early in the causal sequence having a long induction period and those late in the causal sequence a short induction period. The induction period ends when the disease irreversibly begins, and that point to diagnosis is the latent period. Thus, in theory, there is an etiologically relevant time window during which the exposure of interest is causally related to the disease. Further exposure during the induction and latent periods does not contribute to etiology, and including exposure before or after the etiologically important time window, leads to exposure measurement error.

**Selecting the reference period and frequency of data collection**

The limits of knowledge about the biology of most diseases precludes a precise definition of the etiologically relevant time period. Nonetheless, consideration of this concept can provide insight into the appropriate time interval during which to collect exposure data (middle line of figure 1), and the appropriate time period over which the exposure dose variable should be calculated (the *reference period*) (bottom line of figure 1). In practice, for each exposure variable to be calculated (e.g., cumulative dose of hormone replacement therapy), a reference time period needs to be specified. This reference time period is selected based on what is known about the natural history of the disease, i.e., it should reflect the hypothesized etiologically relevant time period. The reference period could be lifetime (up to diagnosis), a physiologic time period (e.g., exposure during adolescence), or could be measured backward from the time of the diagnosis of disease (e.g., the 20 year period ending 5 years before diagnosis). This latter type of reference period is easier to implement in case-control studies, because the date of diagnosis of cases is known at the time of data collection, while in cohort studies, the date of diagnosis is only known after the (prediagnostic) exposure information has been collected. For exposures with long induction and/or latent periods, if the etiologically important time period does not overlap with the study time period, then the best measure of exposure will be retrospective exposure information collected at baseline only. In these situations, the reference period, by necessity, is expressed in relation to baseline, e.g., cumulative dose of aspirin over the 10 years before baseline. For other exposures, the retrospective information from baseline would need to be updated with the prospective information from follow-up contacts to capture exposure during the etiologically relevant time period. For example, to compute hormone replacement therapy use for the 10-year period ending, say, 3 years prior to diagnosis of disease, one would need to com-

---

**FIGURE 1.** Timeline showing the relation between causal exposure-disease sequence, timeline of a cohort study, and reference period for calculation of exposure dose. *Example for a subject who developed the disease endpoint. †Hashmarks equal episodes of exposure.
bined information from baseline with that from, say, yearly updates ending at 3 years before diagnosis (as shown in figure 1). Or, if the exposure acts late in the causal sequence, e.g., aspirin or other anticoagulant drugs in relation to myocardial infarction, then the most recent exposure update(s) before the disease event could be the best exposure measure.

Consideration of the etiologically relevant time period helps answer questions about collection of current versus past exposure data, the frequency of prospective data collection, and the reference period to be used in constructing the dose variable. It is important to note that it is the natural history of the exposure-disease relation that provides answers to these questions, not the “natural history” of the cohort study design. Thus, the advantages of cohort studies, in particular the ability to accurately measure current exposures (e.g., current physical activity, current blood measures) and to measure them prospectively over time are only advantages if they more accurately capture the true (etiologic) exposure. While valid measurement of current physical activity may seem to be an advantage compared with poorly recalled adult lifetime activity, it is quite possible that recalled adult activity is a better measure of the true exposure (e.g., actual lifetime adult activity) than is current activity. Similarly, annual prospective data collection for, say, 10 years would only improve exposure measurement if these years included the etiologically important time window.

Of course, the choice of the frequency of data collection often depends on other study requirements in addition to capturing exposure dose over the reference period. One must consider the frequency needed to obtain accurate endpoint data (if endpoint data are ascertained from the subjects). Frequent (e.g., yearly) contact would usually be needed in order to not lose endpoint information for conditions that lead to major morbidity and mortality. Frequent contact is also needed to keep up-to-date information that allows tracking of subjects’ whereabouts and may help to keep subjects feeling committed to the research project (see the article in this issue by Hunt and White (59)).

Dose and time considerations when there are multiple disease endpoints

One challenge of cohort studies is that there may be multiple disease endpoints, and the appropriate active agent, dose construct, or etiologically relevant time window for a type of exposure could differ for different diseases. For example, the type of physical activity that influences cardiovascular disease risk may differ from the nature of activity that influences colon cancer risk. Colon cancer might be influenced by even low levels of intensity (which might reduce bowel transit time), while a physical activity measure for cardiovascular disease may need to include only the amount of activity of sufficient intensity to have a cardiovascular effect. Similarly, recent use of hormone replacement therapy primarily affects endometrial cancer risk, while only long-term use may affect breast cancer risk. For cohort studies with multiple disease endpoints, instruments must incorporate the items necessary to compute the different exposure dose variables and/or the different etiologically relevant time windows for the various disease outcomes.

Changing the instrument over time

As a cohort study continues over time, new exposures may be added at exposure updates. For example, a food frequency questionnaire was added to the Nurses’ Health Study in year 4, and has been used to assess diet-disease relations for subsequent disease events (15).

A greater challenge in prospective cohort studies is that there may be a need to change the measurement of a specific exposure over time (e.g., one might need to change the way in which dietary fat intake is estimated). While in principle one should never change the measure of an exposure during a study, there are situations for which this may be necessary. First, the available sources of the exposure can change (e.g., new types of hormone replacement therapy have continually been introduced over the last 20 years) or the content of existing products can change (e.g., the fat content of beef has declined over the last several decades). When the external environment or “market place” changes, changes in the questionnaire are needed to continue to accurately measure the exposure. These changes do not necessarily improve the accuracy of the exposure measure; rather, not making the changes would decrease the accuracy. Therefore, such changes should be made to the instrument and to the conversion tables (e.g., fat content of foods) used in the algorithm to create the exposure dose variable. Creating and interpreting a cumulative dose over the reference time period should not be hindered by these types of changes to the instrument. For example, at the exposure updates there may be more types of hormone replacement therapy on the questionnaire which enter into calculation of, say, “cumulative years of hormone replacement therapy use,” but this would not change the interpretation of the exposure variable.

The other type of change in exposure measurement over time would be a change to improve the accuracy of the measure. One way that this happens is that new technology leads to better measures (e.g., a better measure of bone mineral density becomes available).
Another way this occurs, unfortunately, is that the researcher discovers errors or omissions on the original instrument. The decision about whether to change to a more precise measure depends on several factors. First, if one can redo all past measures from stored data or stored specimens using the newer instrument or algorithm, then this should be done. For example, if blood has been stored, it can be reanalyzed using a new method even for older specimens. Or, if a new table for β-carotene in foods becomes available due to better techniques for measuring β-carotene, then one can use the originally collected food frequency questionnaire data and recreate the exposure variable for dietary intake of β-carotene. For situations such as these, there is no reason to keep an older, less accurate exposure measure.

It is most challenging to decide whether to change an exposure measure when it will mean that, for some subjects, or at some data collection time points, the exposure assessment is more accurate than others. In the Framingham study, for example, the measure of serum cholesterol became more precise by the second examination, and questions on cigarette smoking were modified over time to improve accuracy (6). These changes meant that, for example, serum cholesterol measured at baseline was less accurate than that measured at the year 2 clinic visit. One way to deal with two versions of a measure is to drop the earlier, less accurate measure. If the exposure measurement changed at, say, year 4, one could only use the new instrument and only the disease events that occur after year 4. This, however, would reduce study power by loss of the early endpoints. Little has been written about incorporating the use of two different methods for the same exposure in a study. One approach is to stratify the analysis by the instrument used; then one would expect that the relative risk based on the more accurate exposure instrument would be less attenuated due to measurement error than based on the less accurate instrument. This approach has been suggested in the context of case-control studies which must use proxy respondents for some subjects (60, 61).

When the instrument has changed in a cohort study, it is important to not induce differential error by using the two instruments differently for diseased and control subjects. Specifically, if information from the earlier instrument is to be used for the early cases (e.g., those who develop the disease before the new instrument was used), or if information from the earlier instrument is combined with the improved version (e.g., to compute cumulative dose), the analytic techniques must be such that exposures from those with and without the disease outcome are compared on the same instrument or combination of instruments.
Exposure measurement error and nondifferential measurement error. Exposure assessments (including questionnaire, blood, air sampling, etc.) in cohort studies generally are made closer to the time of the true, etiologic exposure than in case-control studies, which should improve the validity of the exposures measured in cohort studies compared with case-control studies. Cohort studies often include repeated exposure measures, which allows for a more accurate measure of exposure during different time periods. Moreover, cohort studies are much less prone to differential measurement error because, in most cohort designs, exposures are measured before disease occurrence. This reduces the likelihood that knowledge about the disease or the physical and psychologic effects of the disease would lead to different validity of the exposure between those with and without disease. These advantages generally apply to both prospective and retrospective cohort designs.

Despite these advantages, cohort studies are subject to many sources of measurement error. Errors can be introduced during any phase of the study (examples are given in Table 2). As implied above, errors can arise from the design of the instrument (i.e., the questionnaire or laboratory procedure to measure the exposure) through failure to correctly specify the specific causative agent, the etiologic important time window, and the appropriate construct of dose (i.e., the appropriate way to incorporate intensity, frequency, and time period of the exposure) for a given disease outcome. Measurement error also arises from errors or omissions in the protocol for use of the instrument and from poor execution of the protocol during data collection, e.g., failure of subjects to follow instructions on self-administered questionnaires or poor handling of biologic specimens. Errors can also arise from limitations due to inherent subject characteristics; these include memory limitations, social desirability bias, and month-to-month variability in biologic characteristics. Some sources of error may be greater in cohort studies than in case-control studies because of the usual longer time frame of cohort studies. In particu-

**TABLE 2. Sources of measurement error in cohort studies**

| Errors in the selection or design of the instrument to measure the exposure |
| Lack of coverage of all sources of the active agent by the instrument |
| Inclusion of exposures that do not have the actual active agent |
| Time period assessed by the instrument is not the etiologically important time period |
| Phrasing of questions that lead to misunderstanding or bias |

| Errors or omissions in the protocol for use of the instrument |
| Failure to specify the protocol (including instructions to the subject) in sufficient detail |
| Failure to specify a method to handle unanticipated situations consistently |

| Poor execution of the study protocol |
| Failure of the data collectors or laboratory technicians to follow the protocol in the same manner for all subjects |
| Failure of the subjects to read or understand the instructions in self-administered questionnaires |
| Improper handling or storage of biologic specimens |

| Limitations due to inherent subject characteristics |
| Memory limitations of subjects including poor recall of past exposures and influence of recent exposures on memory of past exposures |
| Tendency of subjects to overreport socially desirable behaviors (e.g., exercise) and underreport socially undesirable or sensitive behaviors (e.g., induced abortion) |
| Short-term (month to month) variability in biologic characteristics |
| Effect of the outcome under study on the exposure measures (e.g., effect of preclinical disease) |

| Drift in accuracy of exposure measures over time |
| Change in instrument over time |
| Interviewer fatigue |
| Failure to standardize laboratory method periodically |
| Degradation of agent in biologic specimens |

| Errors during data processing and in the creation of the exposure variables |
| Data entry errors |
| Errors in conversion tables used to convert subject responses to units of active agent (e.g., β-carotene by type of food) |
| Development of an exposure variable algorithm that does not express the etiologically important dose or time period |

* Adapted from Armstrong et al. (1).
lar, the accuracy of the instrument could drift, over time, due to failure to standardize the laboratory instrument, fatigue among interviewers, or degradation of the active agent in biologic specimens. Finally, errors enter at the time of data analysis, including key entry errors and errors in conversion tables used to convert subjects' responses to units of the active agent (e.g., errors in the nutrient database used to convert foods to international units of vitamin A).

In cohort studies, most of the sources of error described above would lead to nondifferential error; nonetheless, these and others can be sources of differential error. Major sources would be the effects of the biologic changes during the prediagnostic phase of disease on biologic measures (e.g., undetected bleeding may lead to lower iron status during the prediagnostic phase of colon cancer), the influences of prediagnostic symptoms on a subject's behavior (e.g., gastrointestinal symptoms may lead to reduced fiber intake before diagnosis of colon cancer), and the effects of the subject's risk of the disease on accuracy of reporting (e.g., those with strong family history of colon cancer may be both more accurate in reporting family history and at higher risk of colon cancer).

METHODS FOR REDUCING MEASUREMENT ERROR IN COHORT STUDIES

The primary method to reduce measurement error is the careful selection or construction of the exposure instrument; several issues related to the design of accurate instruments have been reviewed above. Other approaches to reduce measurement error include judicious selection of a study population to improve measurements, use of repeated measures of exposure, and rigorous quality control procedures throughout the conduct of the study. Some steps also can be taken at the time of analysis or interpretation of study results—these include removing outcome events that occur soon after the measurements were taken and the use of information on the accuracy of the exposure measure to "correct" the results. Each of these topics is discussed in more detail below, with emphasis on those issues particularly relevant to cohort studies.

Selection of a cohort to reduce measurement error

In cohort studies, the investigator often has more options in selection of a study group(s) than in case-control studies, which are usually population-based. Several researchers have taken advantage of this by selecting a study cohort that could lead to reduced measurement error. The Nurses' Health Study is an excellent example of a study in which a specific cohort was selected because the cohort members were expected to provide more accurate information on exposures and outcomes than the general population (15, 63). Female registered nurses were selected to be studied because they were assumed to be excellent reporters of exogenous hormone use, which was the original study exposure, and disease outcomes.

Several authors have proposed another approach by which selection of the cohort(s) would reduce measurement error. By selecting a population with larger variation of exposure, the effects of measurement error are reduced (64). For example, the European Prospective Investigation into Cancer and Nutrition (EPIC) includes a multinational population which would have a larger variance of nutrient intake than within a single country (65). If the variance of the (true) exposure is greater in one study population than another, and the variance of the exposure measurement error is the same, then the effect of the measurement error on the relative risk is less in the population with the greater variability of exposure (64). As noted above, the validity coefficient varies inversely with the ratio of the exposure error variance to the true exposure variance. Thus, the validity of the exposure measure can be increased by increasing the true exposure variance as well as by the usual methods to reduce the error variance. This approach of selecting a cohort with larger exposure variance would lead to less attenuation of the relative risk from the effects of nondifferential measurement error. It could also substantially reduce the sample size (or equivalently increase power for a fixed sample size) unless the study population with the larger exposure variance has a significantly lower disease incidence rate than the alternative choices with smaller exposure variance (64).

At least one of the current large cohort studies has attempted to artificially create a cohort with a larger exposure variance through sampling. The American Association of Retired Persons (AARP) Cohort Study, being conducted by the National Cancer Institute, is screening potential cohort members by use of a food frequency questionnaire (66). Subjects are then selected for cohort entry by sampling more heavily from those with the more extreme values, e.g., those in the highest and lowest 10 percent of fat intake. This approach would lead to larger exposure variance than random sampling of this population.

Use of repeated measurements to reduce measurement error

Although issues relating to frequency of data collection were discussed above, another related consideration in determining the frequency of exposure measurement is the advantage of averaging (or summing) two or more measures of the exposure for each sub-
ject. The approach of averaging repeated measures has been recommended as an effective method of decreasing the measurement error, compared with the use of a single measurement (1, 2, 67–69). The measures are typically repeated administrations of the same instrument over time. Use of repeated measures is particularly applicable to prospective cohort studies, because the measures can be averaged over long time periods (e.g., years). For example, the New York University Women's Health Study measured blood hormone concentrations by use of samples collected annually over 5 years (23). On the other hand, the average exposure only increases the validity of the exposure measure if the average over that time period is a good expression of the true exposure.

Fleiss (68) and others (70–72) have shown the degree of improvement in the validity of an exposure variable that results from the combination of multiple measures when the errors in the two or more measures to be averaged are equal and uncorrelated (the parallel test model). "Equal error" means that each measure is an equally precise measure of the true exposure, as would be expected when the same instrument was used for the repeated measures. "Uncorrelated error" means that the measurement error on one administration of the instrument would not be correlated with the error on the second administration, i.e., those subjects whose first measure overestimated their true exposure would be equally likely to have the second measure over- or underestimate their true exposure (beyond any systematic bias in the instrument that affects all subjects). Often repeated laboratory measures based on biologic samples collected at multiple times for each subject (not repeated analysis of the same biologic sample) can be assumed to meet the criteria of the parallel test model, if the true exposure is thought to be the true average over the time period of specimen collection.

When each individual in a study population is measured \( k \) times using repeated parallel measures of exposure variable \( X \), the improvement in validity by use of the average measure, \( A \), for each individual (versus use of only one measure of \( X \)) can be calculated. The correlation of the average with the true exposure, \( \rho_{PA} \) (the validity coefficient of \( A \)), is greater than the correlation of a single measure \( X \) with the true exposure, \( \rho_{TX} \) (the validity coefficient of \( X \)), by

\[
\rho_{PA} = \sqrt{\frac{k \rho_{TX}^2}{1 + (k-1)\rho_{TX}^2}}. \tag{2}
\]

If \( \rho_{TX} \) is known from a validity study, then the validity coefficient for \( A \) can be calculated from the above equation. Table 3 gives examples of the improvement in validity one can achieve by using multiple parallel measures. For example, averaging three measures each with a validity coefficient of 0.5 can yield a new measure with validity coefficient of 0.71, or averaging two measures each with a validity coefficient of 0.85 can yield a new exposure measure with a validity coefficient of 0.92.

Often one does not have a measure of the validity coefficient of \( X \), \( \rho_{TX} \), but, rather, a measure of the reliability of repeated measures of \( X \). \( \rho_{XX} \). \( \rho_{XX} \) may be estimated by the Pearson correlation of two parallel measures \( X \) and \( X' \) (or the average Pearson correlation between pairs of measures, if \( X \) is repeated multiple times) (70, 71). Then the reliability of parallel measures of \( X \), \( \rho_{XX} \), can be used to estimate \( \rho_{TX} \) (70, 71):

\[
\rho_{TX} = \sqrt{\rho_{XX}}. \tag{3}
\]

and \( \rho_{XX} \) can be substituted for in equation 2. Table 3 also gives the improvement in the validity of a measure by use of multiple measures based on values of \( \rho_{XX} \). For example, parallel measures with a reliability coefficient of 0.36 suggests that the validity coefficient of one measure is 0.6 and of the average of two measures would be 0.73.

<table>
<thead>
<tr>
<th>( \rho_{TX} )</th>
<th>( \rho_{XX} )</th>
<th>( k=2 )</th>
<th>( k=3 )</th>
<th>( k=4 )</th>
<th>( k=5 )</th>
<th>( k=10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.30</td>
<td>0.09</td>
<td>0.41</td>
<td>0.48</td>
<td>0.53</td>
<td>0.57</td>
<td>0.70</td>
</tr>
<tr>
<td>0.40</td>
<td>0.16</td>
<td>0.53</td>
<td>0.60</td>
<td>0.66</td>
<td>0.70</td>
<td>0.81</td>
</tr>
<tr>
<td>0.50</td>
<td>0.25</td>
<td>0.63</td>
<td>0.71</td>
<td>0.76</td>
<td>0.79</td>
<td>0.88</td>
</tr>
<tr>
<td>0.60</td>
<td>0.36</td>
<td>0.73</td>
<td>0.79</td>
<td>0.83</td>
<td>0.86</td>
<td>0.92</td>
</tr>
<tr>
<td>0.70</td>
<td>0.49</td>
<td>0.81</td>
<td>0.86</td>
<td>0.89</td>
<td>0.91</td>
<td>0.95</td>
</tr>
<tr>
<td>0.80</td>
<td>0.64</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>0.85</td>
<td>0.72</td>
<td>0.92</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
<td>0.98</td>
</tr>
<tr>
<td>0.90</td>
<td>0.81</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
</tr>
</tbody>
</table>

* \( \rho_{XX} \) is the correlation coefficient between each (parallel) exposure measure, \( X \), and the true exposure, \( T \). \( \rho_{XX} \) is the correlation between any two (parallel) exposure measures, \( X \) and \( X' \) (\( \rho_{XX} = \rho_{XX} \)). \( \rho_{TX} \) is the correlation of \( A \), the average of \( k \) (parallel) exposure measures, and \( T \), the true exposure.
The primary advantage of averaging two or more parallel measures is that the increase in the validity of the exposure measure will reduce the degree of attenuation of the relative risk due to (nondifferential) measurement error (equation 1), and, therefore, reduce the required sample size (68).

One disadvantage of the use of repeated measures is that the cost per subject would increase due to the repeated study contacts. Thus the total study cost could be increased or decreased depending on the trade-off between the reduced sample size and the increased cost per subject. One way to select \( k \), the number of repeated measures per subject, is to determine \( k \) which would minimize total costs (68).

The possibility of violation of the assumptions of the parallel test model must be considered when evaluating the use of multiple measures (1, 2). It may be that the two or more measures to be averaged are not equally precise. In this situation, the average of a precise measure with a less precise measure may result in a measure which is less valid than the better measure alone. An obvious example is the use of the average of a perfect measure of exposure and an imperfect measure.

When the two or more measures to be averaged have correlated errors, but the other assumptions of parallel tests hold, the improvement in the validity of the exposure measure will be less than that predicted by equation 2 or table 3. If the errors are perfectly correlated (i.e., the measure is perfectly repeatable, even though it is not a perfect measure of exposure), the use of multiple measures will lead to no improvement in validity. Most questionnaire measures would have some degree of correlated error. For example, subjects whose fat intake was underestimated on the administration of a food frequency questionnaire would tend to have their fat intake underestimated on the second administration (due to some subjects consistently underreporting high fat foods, failure of the instrument to include certain high fat foods frequently eaten by some subjects, etc.). Similarly, repeated analysis of the same biologic specimen would have correlated error, e.g., if a subject's serum cholesterol was higher than her/his long-term average on the day the blood was drawn, it would tend to be higher on each repeated analysis of the sample.

Finally, when the repeated exposure measurements are not collected uniformly over time, a simple average might not be appropriate. In the Port Pirie Study, blood lead measures were taken at birth, 6, 15, and 24 months, and then annually. A composite, lifetime measure was computed for each subject as the area under the time-exposure curve (17,18).

**Quality control**

Quality control procedures are an important component of exposure measurement, but can only be briefly mentioned here. When studies are large, as cohort studies often are, it is even more essential to implement systematic quality control procedures to reduce measurement error. It is important to have a complete and detailed study procedures manual, to pretest instruments well, to have standardized training and monitoring of data collectors, and to edit data as they are collected (1, 73). These and other quality control issues are covered in the article in this issue by Whitney (74).

**Use of information from validity and reliability studies**

The conduct of validity or reliability studies on a subsample of the cohort has become an important and common component of many large cohort studies (75–82). Some cohort studies have incorporated a validity substudy in which the measured exposure using the cohort study instrument is compared with a near perfect measure of exposure which is too expensive or too burdensome to be used in the full study (75). Since perfectly valid exposure measures are not often available, validity studies are usually not feasible. More often test-retest reliability studies are conducted in which repeated applications of the exposure instrument are compared, or relative validity or intermethod reliability studies are conducted in which the measured exposure is compared with a more accurate (but less than perfect) comparison measure (76–82). If the comparison measure is carefully selected so that certain assumptions are met, these types of studies can also provide some information on the validity of the measure (1, 70, 71). The design, analysis, and interpretation of validity and reliability studies is beyond the scope of this presentation, but is covered by others (1, 44, 68, 70, 72, 83, 84).

One goal for the use of the information from validity and reliability substudies in cohort studies is to estimate the effects of the inaccuracy of the exposure measure on the relative risk (38, 69). Development of statistical techniques which incorporate the information from validation substudies to correct relative risk estimates for the effects of measurement error is an active area of methodological research (85–92). Work is also currently being conducted on the optimal design of the validation substudy (93).

**Reducing measurement error at the time of analysis by dropping early events**

As noted above, one source of differential measurement error is the effect of preclinical disease on the
exposures under study. In case-control studies, the general approach to removing the effects of symptoms or preclinical disease on exposures is to have the reference period end at a fixed time before diagnosis, e.g., fiber intake could be assessed for a 10 year reference period ending 2 years prior to diagnosis, and for a comparable time for controls. Although one could attempt to simulate this approach in cohort studies, another common approach to remove this source of differential error is to drop from analysis cases who develop the disease within a certain time after the exposure was measured. For example, in a cohort study of colon cancer and serum cholesterol, removing the cancer cases that occurred within the first several years after cholesterol measurement changed the relative risk from less than 1 to close to 1 (94). Thus, what appeared to be a protective effect of serum cholesterol on cancer risk was more likely an effect of preclinical colon cancer in reducing serum cholesterol.

Avoiding inducing differential measurement error during data analysis

Care must be taken to select the appropriate reference period and the appropriate statistical method so that the exposure measures for those with and without the disease outcome are computed in a similar manner. This is a particular concern in cohort studies, because exposure assessment ends at the outcome event for those who develop the disease but continues to end-of-study for most participants who do not develop the disease. The control reference period needs to be similar in length and measured over the same chronologic years as the case reference period to avoid inducing differential measurement error.

SUMMARY

Cohort study designs have several advantages over case-control studies in terms of exposure measurement. If exposure measurement occurs before disease occurrence, cohort studies are much less prone to differential measurement error. Prospective data collection should also reduce measurement error due to poor recall of past exposures. The primary drawback of cohort studies is the large sample size leading to high data collection costs. Several approaches to reduce such costs have been discussed in this presentation, such as selection of lower cost measurement methods and fully measuring the exposure only on a subsample of the cohort (e.g., nested case-control design). However, other innovative approaches to reduce costs are needed. In addition, study reviewers should also consider that the higher costs are justified in relation to the several benefits of this study design, which include not only less measurement error, but also less susceptibility to selection bias and often the ability to study multiple disease outcomes.

Improving the accuracy of exposure measurement is increasingly important for cohort studies as we move on to the study of exposures that are difficult to measure or to those with lower relative risks of disease. In such studies, attenuation of the relative risk by the effects of measurement error can lead to failure to detect an association between exposure and disease. The validity of exposure measurements could be improved by a better understanding of the biologically active agent and etiologically important time period of the exposure-disease relation, and by incorporating these into the measure. Long-term cohort studies which cover the etiologically relevant time period could improve the accuracy of measures of exposures by use of repeated biologic measures or repeated updates of self-reported exposures. Measurement error also can be reduced by judicious choice of a cohort to study and by careful attention to quality control procedures. Continued emphasis on the evaluation and improvement of the measurement properties of instruments used in epidemiologic studies will improve the validity of the results of cohort studies.

REFERENCES


