10. Sources of error

A systematic framework for identifying potential sources and impact of distortion in observational studies, with approaches to maintaining validity

We have already considered many sources of error in epidemiologic studies: selective survival, selective recall, incorrect classification of subjects with regard to their disease and/or exposure status. Because of the limited opportunity for experimental controls, error, particularly "bias", is an overriding concern of epidemiologists (and of our critics!) as well as the principal basis for doubting or disputing the results of epidemiologic investigations.

Accuracy is a general term denoting the absence of error of all kinds. In one modern conceptual framework (Rothman and Greenland), the overall goal of an epidemiologic study is accuracy in measurement of a parameter, such as the IDR relating an exposure to an outcome. Sources of error in measurement are classified as either random or systematic (Rothman, p. 78).

Rothman defines ***random error*** as "that part of our experience that we cannot predict" (p. 78). From a statistical perspective, random error can also be conceptualized as sampling variability. Even when a formal sampling procedure is not involved, as in, for example, a single measurement of blood pressure on one individual, the single measurement can be regarded as an observation from the set of all possible values for that individual or as an observation of the true value plus an observation from a random process representing instrument and situational factors. The inverse of random error is ***precision***, which is therefore a desirable attribute of measurement and estimation.

Systematic error, or ***bias***, is a difference between an observed value and the true value due to all causes other than sampling variability (Mausner and Bahn, 1st ed., p. 139). Systematic error can arise from innumerable sources, including factors involved in the choice or recruitment of a study population and factors involved in the definition and measurement of study variables. The inverse of bias is validity, also a desirable attribute.

These various terms – "systematic error", "bias", "validity" – are used by various disciplines and in various contexts, with similar but not identical meanings. In statistics, "bias" refers to the difference between the average value of an estimator, computed over multiple random samples, and the true value of the parameter which it seeks to estimate. In psychometrics, "validity" most often refers to the degree to which a measurement instrument measures the construct that it is supposed to measure. The distinction between random and systematic error is found in many disciplines, but as we shall see these two types of error are not wholly separate. We will return to the issue of terminology below, in the section on "Concepts and terminology".)

Precision

The presence of random variation must always be kept in mind in designing studies and in interpreting data. Generally speaking, small numbers lead to imprecise estimates. Therefore, small differences based on small numbers must be regarded with caution since these differences are as likely the product of random variation as of something interpretable.

Estimates of ratio measures (e.g., the relative risk) based on sparse data are very susceptible to instability. For example, a relative risk of 5.0 based on the occurrence of 4 cases among the nonexposed becomes twice as large (10.0) if two unexposed cases are missed through the vagaries of sampling, measurement, missing data, or other reasons. If three are missed, the relative risk will be 20.

Example to illustrate the concept of precision

Consider the following data, from Table 4 of Hulka et al., "Alternative" controls in a case-control study of endometrial cancer and exogenous estrogen, *Am J Epidemiol* 112:376-387, 1980:

Effect of duration of estrogen use on relative risks [age-adjusted] using three control groups among white women, North Carolina, 1970-76

Duration of use	No. of Cases	D&C Controls		Gynecol. Controls		Community Controls	
		No.	RR	No.	RR	No.	RR
None used	125	136		118		172	
Less than 6 months	8	13	0.7	12	0.7	20	0.8
6 months - 3.5 yrs.	9	14	0.7	9	0.9	21	0.7
3.5 yrs. - 6.5 yrs.	9	16	0.8	1		7	1.7
6.5 yrs. - 9.5 yrs.	9	11	1.2	2	3.8	5	2.5
More than 9.5 yrs	19	10	2.0	2	5.1	4	5.5

Note: "D&C Controls" = dilatation and curetage patients as controls
"Gyn Controls" = other gynecology clinic patients as controls

First, let's recall from our previous topic that since these data come from a case-control study, the "relative risks" in the table are odds ratios. Since the disease is a rare one, however, odds ratios, risk ratios, and incidence density ratios will all be about the same. Also from that lesson we should be able to reformulate the above data as a series of 2 x 2 tables if for some reason we wished to. Such reformulation would make it easier to see how to calculate crude relative risk estimates (OR's) from

the data in the table. (Note that the OR's in the table are age-adjusted, through a mathematical modeling procedure called multiple logistic regression, so our crude OR's will differ in some cases.)

Notice that the relative risk estimates for the GYN and community controls in the longer duration categories are based on very few controls. For example, if two of the community controls classified as duration of use "3.5 years - 6.5 years" were instead duration "6.5 years - 9.5 years", then the relative risk estimates would be reversed and the consistent dose-response picture would no longer appear. [For the moment we are ignoring the age adjustment, though for these two particular duration groups in the community controls the adjusted OR's are the same as if they were calculated from the data in the table.] Similarly, the RR's over 5 for the longest duration subjects are based on 2 and 4 controls in the GYN and community groups, respectively. On the other hand, the fact that similar results were found in two control groups strengthens the assessment that a dose-response relationship truly exists, rather than being a chance finding.

Quantifying the degree of precision or imprecision – confidence intervals

Statistical techniques such as standard errors and confidence inteverals are used to quantify the degree of precision or imprecision of estimates; there are also rules of thumb (e.g., see Alvan Feinstein, *J Chron Dis* 1987; 40:189-192). A ***confidence interval*** provides a range of values that is expected to include the true value of the parameter being estimated. The narrower the confidence interval, the more precise the estimate.

For example, suppose we are estimating the relative risk for endometrial cancer in women who have used replacement estrogens for 3.5 years to 6.5 years (compared to women who have not taken estrogens) and that the "true" (but unknown) relative risk is 1.5. Suppose also that the estimate we obtain from the data for community controls in the above table are unbiased, though they do reflect random error. Hulka et al. computed the age-adjusted value as 1.7 (the crude value is very similar: 1.77). The 1.7 is a point estimate and provides our best single estimate of the (unknown) true relative risk of 1.5.

We do not expect that our estimate is exactly correct, however, so we also compute an interval estimate, or confidence interval as an indicator of how much data were available for the estimate. Suppose that the 95% confidence interval is (0.6,4.7). The interpretation would be that 1.7 is the best single estimate of the (unknown) true relative risk and that there is "95% confidence that the true relative risk is somewhere between 0.6 and 4.7". "Confidence" does not mean the same thing as "probability". In this case "95% confidence" means that we obtained the confidence limits 0.6 and 4.7 through a procedure that yields an interval that will contain the true value in 95% of instances in which we use it and will not contain the true value in the remaining 5% of instances.. Loosely speaking, a 95% confidence interval of 0.6-4.7 means that the observed value of 1.7 is "compatible", by conventional usage, with true relative risks anywhere between 0.6 and 4.7 inclusive.

Another way of describing the meaning of "compatible" is the following. The limits 0.6 and 4.7 are obtained from the point estimate of 1.7 and the estimated standard error of that estimate. The estimated standard error is a function of the size of the numbers (i.e., the amount of data) on which

the point estimate is based and thus a measure of its imprecision. A 95% confidence interval (0.6,4.7) means that if our study had yielded a point estimate anywhere in that interval, the 95% confidence interval around that point estimate would contain the value 1.7. In that sense the observed value of 1.7 is compatible with true relative risks anywhere between 0.6 and 4.7.]

In fact, the original table in Hulka et al. did include confidence intervals for the odds ratios. Attention to confidence intervals or to sparseness of data is an important aspect of interpreting results.

Reducing random variation (increasing precision)

Confidence intervals and other procedures for assessing the potential for random variation in a study do not increase precision, but merely quantify it. The major strategies for reducing the role of random error are:

1. Increase sample size – a larger sample, other things being equal, will yield more precise estimates of population parameters;
2. Improve sampling procedures – a more refined sampling strategy, e.g., stratified random sampling combined with the appropriate analytic techniques can often reduce sampling variability compared to simple random sampling;
3. Reduce measurement variability by using strict measurement protocols, better instrumentation, or averages of multiple measurements.
4. Use more statistically efficient analytic methods – statistical procedures vary in their efficiency, i.e., in the degree of precision obtainable from a given sample size;

Bias

Bias is by definition not affected by sample size. Rather, bias depends on enrollment and retention of study participants and on measurement. [A technical definition of "bias" in its epidemiologic usage (based on Kleinbaum, Kupper, and Morgenstern) is the extent to which an estimate differs from the true value of the parameter being estimated, even after sample size is increased to the point where random variation is negligible. This definition is based on the statistical concept of consistency; in statistics, an estimator is ***consistent*** if its value continuously approaches the value of the parameter it estimates as the sample size increases.]

Concepts and terminology

In the area of bias and validity, as in so many other areas that cross disciplines, terminology can be a significant source of confusion. Such dangers are particularly apparent when the terms are also used in a nontechnical sense in ordinary discourse. An additional source of confusion for terminology concerning validity is overlap among the concepts. For example, measurements are an ingredient of studies, but studies can also be regarded as measurement procedures applied to populations or

associations. So the same terms may be applied to individual measurements and to entire studies, though the meaning changes with the context.

Internal validity

Epidemiologists distinguish between internal validity and external validity. ***Internal validity*** refers to absence of systematic error that causes the study findings (parameter estimates) to differ from the true values as defined in the study objectives. Systematic error can result from inaccurate measurements of study variables, nonuniform recruitment or retention of study participants, or comparisons of groups that differ in unknown but important characteristics. Thus, internal validity concerns bias in estimates for the target population specified in the study objectives.

External validity

External validity refers to the extent to which a study's findings apply to populations other than the one that was being investigate. Generalizability to populations beyond the target population for which the study was designed and/or beyond the circumstances implicit in the study is a matter of scientific inference, rather than a technical or statistical question (see Rothman and Greenland). Therefore external validity is probably better considered in relation to causal inference and interpretation of study results. (Rothman and Greenland regard "external validity" as a misnomer, preferring to draw the distinction between validity and generalizability.)

Validity pertains to a specific measure

Since different types of errors affect specific findings in different ways, validity must generally be discussed in regard to a specific measure or measures. A study aimed at testing an etiologic hypothesis typically seeks to estimate strength of association measured as the ratio or difference of incidences in different groups. Lack of internal validity in this context means inaccuracy (bias) in these estimates. In fact, Kleinbaum, Kupper, and Morgenstern (*Epidemiologic Research*, ch. 10) define (internal) validity and bias in terms of systematic distortion in the "measure of effect". A study can yield a valid (unbiased) measure of effect despite systematic errors in the data if the errors happen to offset one another in respect to the measure of effect. However, a study with no systematic error can yield a biased estimate of a measure of effect (for example, due to random variability in an important measurement – see appendix). Much of the methodologic writing about bias concerns distortion in effect measures.

Not all studies have as their objective the estimation of a measure of effect, and even studies that do also report estimates of other parameters (e.g., incidence rates, prevalences, means). Thus, even if the measure of effect is accurately estimated, the possibility and extent of bias in other measures must be considered.

Measurement validity

In Rothman and Greenland's perspective, measurement is the purpose of all studies, so the concept of validity of measurement is the same as that of validity. However, validity of the ***measurements*** carried out in conducting a study raises issues of its own and is addressed in another category of

methodological literature. Validity of measurement (I have to confess that this is my own term to differentiate this type of validity) concerns the avoidance of error in measuring or detecting a factor (e.g. blood pressure, smoking rate, alcoholism, HIV infection). The sociologic and psychologic literature deals extensively with measurement validity, particularly in relation to data collected via questionnaires and interviews. Cognitive psychology studies the thinking processes by which study participants decode questionnaire items and retrieve the information from memory (e.g., Warnecke *et al.*, 1997a; Warnecke *et al.*, 1997b). Psychometrics studies statistical aspects of psychological measurement instruments (Nunnally, 1994). These disciplines are especially pertinent for epidemiologists interested in sophisticated measurement of self-report measures.

Direction of bias – "which way is up"

Concepts and terminology can also complicate descriptions of the direction in which a bias may distort a measure of effect. The sources of confusion are: (1) an association can be positive (RR > 1.0) or inverse (RR < 1.0, also referred to as "negative"), (2) a source of bias can make a measure of effect increase in magnitude, decrease in magnitude, move towards 1.0 from either above or below, and move away from 1.0 in either direction, and (3) it is easy to lose sight of whether the measurement of association being referred to is that observed in the study or the "true" one that exists in the target population. [Try plotting some relative risks on a line as you read the next two paragraphs.]

Describing the direction of bias – example:

Suppose that "aggressive" people are more likely to survive an acute myocardial infarction (MI) than are nonaggressive people. A case-control study of MI that recruits its cases from among (live) hospitalized MI patients will therefore overrepresent aggressive MI cases, since proportionately more of them will live long enough to enroll in the study. If this is the only source of systematic error, then we expect the observed relative risk (RR) to be greater than the true relative risk for incidence of acute MI (since the true relative risk would include the victims who died before they could be enrolled in the study). The direction of bias is in the positive direction (toward higher values of the RR), regardless of whether the true RR is greater than 1.0 (i.e., aggressive people also more likely to have an MI) or less than 1.0 (aggressive people are less likely to have an MI).

In contrast, uniform random error in the measurement of aggressiveness independent of other variables typically moves the observed RR "toward the null" (closer to 1.0 than the true RR). Bias toward the null can produce a lower observed RR (if the true RR is greater than 1.0) or a higher observed RR (if the true RR is less than 1.0), but not an RR that is farther from the null than the true RR. On the other hand, the bias from greater survival of aggressive MI cases in the above hypothetical case-control study will be closer to 1.0 only if the true RR is less than 1.0 and farther from 1.0 only if the true RR is greater than 1.0.

For these reasons we need four terms to characterize the potential effects of sources of bias:

"***Positive bias***" – The observed measure of effect is a larger number than the true measure of effect is (if it could be known);

"***Negative bias***" – The observed measure of effect is a smaller number than the true measure of effect is (if it could be known);

"***Towards the null***" – The observed measure of effect is closer to 1.0 than the true measure of effect is (if it could be known);

"***Away from the null***" – The observed measure of effect is farther from 1.0 than the true measure of effect is (if it could be known);

Another way of describing the direction of bias is to say that the observed measure of effect overestimates (underestimates) the true measure. With this phraseology, however, more information must be available, since "overestimates" could be taken as meaning higher in numerical value or greater in strength (farther from the null).

In the interests of precise communication, we will try to adhere to the above usage, which does not appear to be standard in the profession. However, terminology is only one source of confusion. Consider the longstanding proposition that nondifferential misclassification (covered below) of a dichotomous exposure or disease variable, in the absence of confunding (see next chapter) always produces bias that is "toward the null". This proposition holds as long as nondifferential (independent) misclassification is no worse than what would result from classification each observation by tossing a coin. However, extreme nondifferential misclassification (in the limiting case, misclassification of every participant), however, can bias the measure of effect beyond and then away from the null value.

Types of bias

Students of epidemiology often wish for a catalog of types of bias in order to be able to spot them in published studies. David Sackett (Bias in analytic research. *J Chron Dis* 32:51-63, 1979) once attempted to develop one. Nine sample entries he describes are:

1. Prevalence-incidence (Neyman) bias

   This is Sackett's term for, among other things, selective survival. Also included are the phenomena of reversion to normal of signs of previous clinical events (e.g., "silent" MI's may leave no clear electrocardiographic evidence some time later) and/or risk factor change after a pathophysiologic process has been initiated (e.g., a Type A may change his behavior after an MI), so that studies based on prevalence will produce a distorted picture of what has happened in terms of incidence.

2. Admission rate (Berkson) bias

   Where cases and/or controls are recruited from among hospital patients, the characteristics of both of these groups will be influenced by hospital admission rates.

3. Unmasking (detection signal) bias

   Since by necessity, a disease must be detected in order to be counted, factors that influence

disease detection may be mistakenly thought to influence disease occurrence. This possibility is particularly likely where the disease detection process takes place outside of the study (e.g., in a case-control study), where the disease has an occult, or asymptomatic, phase, and where the exposure leads to symptoms that induce the individual to seek medical attention.

4. Non-respondent bias

   Non-respondents to a survey often differ in important ways from respondents. Similarly, volunteers often differ from non-volunteers, late-respondents from early respondents, and study dropouts from those who complete the study.

5. Membership bias

   Membership in a group may imply a degree of health which differs systematically from others in the general population. For example, the observation that vigorous physical activity protects against CHD was initially thought likely to be a result of fitter people (with lower innate CHD risk) being more likely to engage in vigorous activity. Another example would be if people who participate in a health promotion program subsequently make more beneficial lifestyle changes than nonparticipants due not to the program itself but to the participants' motivation and readiness to change.

6. Diagnostic suspicion bias

   The diagnostic process includes a great deal of room for judgment. If knowledge of the exposure or related factors influences the intensity and outcome of the diagnostic process, then exposed cases have a greater (or lesser) chance of becoming diagnosed, and therefore, counted.

7. Exposure suspicion bias

   Knowledge of disease status may influence the intensity and outcome of a search for exposure to the putative cause.

8. Recall bias

   Recall of cases and controls may differ both in amount and in accuracy (selective recall). Cases may be questioned more intensively than controls.

9. Family information bias

   Within a family, the flow of information about exposures and illnesses is stimulated by, and directed to, a family member who develops the disease. Thus a person who develops rheumatoid arthritis may well be more likely than his or her unaffected siblings to know that a parent has a history of arthritis.

The appendix to Sackett's article gives his entire catalog of biases.

Classifying sources of bias

Inspite of David Sackett's initiative, a complete catalog of biases does not yet exist. Instead, following Olli Miettinen's work in the 1970's, epidemiologists generally refer to three major classes of bias:

1. ***Selection bias*** – distortion that results from the processes by which subjects are selected into the study population:
2. ***Information bias*** (also called ***misclassification bias***) – distortion that results from inaccuracies in the measurement of subject characteristics, and incorrect classification therefrom:
3. ***Confounding bias*** – distortion in the interpretation of findings due to failure to take into account the effects of disease risk factors other than the exposure of interest.

Confounding bias is somewhat different from the other two forms in that the actual data collected by the study may themselves be correct; the problem arises from a misattribution of observed effects (or their absence), i.e., an apparent effect is attributed to the exposure of interest, whereas in fact it ought to have been attributed to some other factor. We will discuss confounding in the following chapter.

Of course, as in so many other areas of epidemiology, the divisions among the classes are only relative, not absolute!

Selection bias

Ignoring the questions of random error in sampling (i.e., assuming that all samples are large enough so that random variation due to sampling is negligible), we can see that if the process by which subjects are recruited favors or overlooks certain types of subjects, then the study population we obtain will not be representative of the population for which we are attempting to obtain estimates. For example, if we are studying characteristics of persons with diabetes and obtain all of our subjects from among hospital patients, the characteristics of this study population will yield a distorted or biased estimate of the characteristics of diabetics in general.

In case-control studies, situations that can produce selection bias include:

- the exposure has some influence on the process of case ascertainment ("detection bias"): the exposure prevalence in cases will be biased;
- selective survival or selective migration – the exposure prevalence in prevalent cases may be biased compared to that in incident cases;
- the exposure has some influence on the process by which controls are selected (e.g., use of chronic bronchitis patients as controls for a study of lung cancer and smoking): the exposure prevalence in controls will differ from that in the base population.

In cohort studies, the primary source of selection bias is generally differential attrition or loss to follow-up. Example (hypothetical):

Complete cohort:

	Type A	Type B	
CHD	40	20	
CHD	160	180	
Total	200	200	RR=2.0

Observed cohort:*

	Type A	Type B	
CHD	32	18	
CHD	144	162	
Total	176	180	RR=1.82

*based on a 10% loss rate among subjects, except that Type A subjects who developed CHD are assumed to have been lost at a 20% rate. If all subjects, including the CHD/Type A group had experienced a 10% loss rate, the incidence in each behavior type group, and therefore the risk ratio, would be undistorted.

Conceptual framework

[After Kleinbaum, Kupper and Morgenstern, *Epidemiologic Research* and *Am J Epidemiol* article on selection bias (see bibliography)].

External population: the population of ultimate interest, but which we are not attempting to study directly – e.g., we may wish to study the relationship between hypertension and stroke in general, but study only subjects in North Carolina, recognizing that generalizing to other areas will require consideration of differences between North Carolina and those other areas. We will not concern ourselves with generalizability in this chapter.

Target population: the population for which we intend to make estimates.

Actual population: the population to which our estimates actually apply. This population may not be obvious or even knowable.

Study population: the group of participants for whom we have collected data. In Kleinbaum, Kupper, and Morgenstern's framework, the study population is regarded as an unbiased sample of the actual population, differing from it only from through unsystematic sampling variability error.

The study population is a subset of the actual population. Bias is the discrepancy between the actual and target populations. Generalizability deals with inference from the target population to an external population (see previous page).

In thinking about selection bias and its potential effect on study results, we find it useful to consider the probabilities according to which people in the target population could gain access to the actual population. These probabilities are called (population) selection probabilities.

For simplicity, consider a dichotomous disease and dichotomous exposure classification, and let the fourfold table in the target population and actual population be as follows:

	E	$\bar{E}$
D	A	B
$\bar{D}$	C	D

Target

	E	$\bar{E}$
D	A^o	B^o
$\bar{D}$	C^o	D^o

Actual

We can then define four selection probabilities:

alpha (α) = (A^o/A) the probability that a person in cell A (in the target population) will be selected into the actual population from which the study population is a random sample

beta (β) = (B^o/B) the probability that a person in cell B (in the target population) will be selected into the actual population

gamma (γ) = (C^o/C) the probability that a person in cell C (in the target population) will be selected into the actual population

delta (δ) = (D^o/D) the probability that a person in cell D (in the target population) will be selected into the actual population

Example: assume that selective survival exists, such that cigarette smokers who suffer an MI are more likely to die before reaching the hospital. Then a case-control study of MI and smoking, using hospitalized MI patients as cases will have alpha lower than beta (exposed cases are less available to study than are nonexposed cases). This bias will produce a distortion in the odds ratio that will understate a true association between smoking and MI (i.e., negative bias).

The assignment for this lecture has an exercise that asks you to apply this conceptual framework to a detection bias issue involving endometrial cancer and estrogen. The basic issue is that use of estrogen might lead to uterine bleeding, which would result in a woman seeking medical attention and receiving a dilation and curettage (D&C). If an occult (asymptomatic) endometrial cancer were

present, then the D&C would detect it. According to the detection bias scenario, women with occult endometrial cancer are therefore more likely to come to medical attention if they are estrogen users, creating a detection bias situation.

This scenario was vigorously disputed, since it depends upon the existence of a sizable reservoir of asymptomatic endometrial cancer, and is now widely discounted. Nevertheless, the endometrial cancer and estrogen issue provides abundant illustrations for concepts related to selection bias and information bias. We will take up this case study presently. (Note that though bias in case-control studies has attracted the most theoretical interest, all study designs are vulnerable.)

Recourse — Minimize loss to follow-up, obtain representative study populations, anticipate sources of bias and avoid them. Sometimes the factors associated with selection bias can be measured, in which case the analysis of the data can attempt to take these factors into account. Logic in the interpretation of the data may be able to marshall evidence for or against selection bias as having been responsible for an observed association. But if you can avoid it, that's the best!

Estrogen and endometrial cancer case example

During the 1970s, case-control studies reported a strong (OR about 10) association between endometrial cancer and use of postmenopausal estrogens. The association was biologically plausible, since the endometrium of the uterus is an estrogen-responsive tissue. Also, endometrial cancer rates were rising in geographical areas where use of postmenopausal estrogens was growing most rapidly.

Criticism of case-control studies had also been rising, however. For one, case-control studies reporting an association between breast cancer and the anti-hypertensive medication reserpine had received wide attention, but the association was later discounted. Also, critics of the case-control design (notably Alvan Feinstein, who labelled the design the "trohoc" study ["cohort" spelled backwards]) had become prominent. The *Journal of Chronic Disease* (now called the *Journal of Clinical Epidemiology*) hosted a conference of leading epidemiologists to discuss the validity of the design (proceedings published in Michel A. Ibrahim and Walter O. Spitzer. *The case-control study: consensus and controversy*. Pergamon, New York, 1979).

At about this time, Barbara Hulka, Carol J.R. Hogue, and the late Bernard G. Greenberg (then all at the UNC School of Public Health) published a comprehensive review of methodological issues involved in the estrogen-endometrial cancer association (Methodologic issues in epidemiologic studies of endometrial cancer and exogenous estrogen. *Amer J Epidemiol* 1978; 107:267-276). The case-control design is particularly susceptible to selection bias, because since the disease has already occurred, the validity of the study is critically dependent upon the selection of cases and controls. The Hulka et al. review made the following points (more material from this review is presented in the appendix to this chapter):

1. Ascertainment of cases

   Cases provide an estimate of estrogen exposure in women who develop endometrial cancer.

This estimate of the prevalence of exposure among cases can be expressed in probability terms as Pr(E|D) – the probability of exposure conditional on having the disease.

Cases in the study should therefore be representative of all similarly described (i.e., age, geography, subdiagnosis) persons who develop the disease with respect to exposure status. For endometrial cancer, two issues are -

a. Heterogeneity of cases (stage, grade, histological type) may reflect different underlying etiology or relationship to exposure.
b. Sources of cases and diagnostic process may have implications for exposure status (e.g., cases from rural hospitals may have had less access to postmenopausal estrogens).

2. Selection of controls

Controls provide an estimate of the prevalence of exposure in the source population from which the cases arose (now referred to as the "study base"). This prevalence can be expressed in probability terms as Pr(E) - the probability that a person selected at random from the study base is exposed to exogenous estrogens. Controls must therefore be representative of the study base with respect to exposure status, so that the prevalence of estrogen use in controls (in probability terms, Pr(E|not D) accurately estimates exposure in the study base. In addition, controls should be able to provide exposure and other data with accuracy equivalent to that obtainable from cases (this point concerns information bias and will be discussed later in this chapter).

Therefore, controls should be similar to cases in terms of:

a. Data sources, so that the opportunity to find out about prior estrogen use is equivalent to that for cases;
b. Other determinants of the disease that cannot be controlled explicitly

But controls should not be too similar to cases on nondeterminants of the disease.

Overmatching and the selection of controls

This last qualification was directed at the issue of detection bias raised by Feinstein (see above). Ralph Horwitz's and Feinstein's recommendation for reducing detection bias was to select controls from among women who had had the same diagnostic procedure as had the cases (dilation and curettage), thereby ensuring that controls did not have occult disease and making them more similar to the cases. Hulka et al.'s response was that such a selection procedure for controls constitutes overmatching.

The concept of overmatching and Horwitz and Feinstein's proposed "alternative controls" (*NEJM*, 1978) focus on the relationship of selection bias and the selection of the control group in a case-control study, which is why the estrogen – endometrial cancer topic is such an excellent one for understanding control selection.

Controls in an experiment

In a true experiment, in which one group is given a treatment and another serves as a control group, the optimum situation is generally for the treatment and control groups to be as close to identical as possible at the time of the treatment and to be subjected to as similar as possible an environment apart from the treatment. If randomization of a large number of participants is not feasible, the control group is matched to the experimental group to achieve as much similarity as possible in anything that might affect development of the outcome.

Earlier generations of epidemiologists were often taught that, by analogy, the control group in a case-control study should be similar to the case group in all characteristics other than disease (and exposure status, which the study seeks to estimate). In that way, exposure differences could more readily be attributed to the effects of the exposure on disease risk, the only other point of difference. Toward that objective, controls have often been matched to cases to increase similarity of the groups.

Analogies between experimental and case-control study designs

However, the analogy between the control group in a case-control study and the control group in an experiment, is faulty. In an experiment, exposure is introduced in one of two hopefully equivalent groups, and outcomes subsequently develop. The control group is chosen to have equivalent risk for the outcome in the absence of the exposure. In a case-control study, exposures exist in a population, and outcomes develop. The equivalence that is required for a valid comparison is that between exposed and unexposed persons. The case group – the members of the population who have developed the outcome – are not located in a corresponding position vis-a-vis the disease process as are the exposed group in a true experiment. The former is a group of people who develop the outcome; the latter are a group at risk for the outcome.

The correct experimental analog to the case group in a case-control study is the group of participants who develop the outcome during the experiment. In both designs, the cases arise from a population of both exposed (or "experimental") and unexposed (or "control") persons. Similarly, the correct analog for the control group in a case-control study is a random sample of all participants in the experiment at some point following the onset of exposure. The set of all participants in the experiment is the "study base" for the experiment. If a case-control study is conducted using the cases which arose in that experiment, then the control group should serve to estimate the proportion of exposure in that study base.

Matching and selection bias

Forcing the control group to be similar to the case group, either through matching or through using a source for recruitment of controls similar to that for recruitment of cases, will ordinarily make the control group less like the study base and may therefore introduce selection bias. Whether or not selection bias will be introduced depends upon the analysis methods used and whether or not the matching factors are related to prevalence of exposure. If the characteristics are unrelated to exposure then selection bias will not occur for that exposure, since both the matched and

unmatched control groups will presumably yield the same estimate of exposure prevalence. If the characteristics are risk factors for the disease, then although matching may introduce selection bias, this bias can be eliminated by controlling for the matching factors in the analysis (think of each matching factor as identifying subsets in both the cases and study base, so that the overall study can be regarded as a set of separate, parallel case-control studies, each itself valid).

Overmatching

However, if the characteristics are related to the exposure and are not risk factors for the disease, then forcing the controls to be more like the cases will distort both the exposure prevalence in controls (making it more like that in the cases and less like that in the study base) and odds ratio relating exposure and disease. This scenario is termed ***overmatching***. If the matching factors are controlled in the analysis (which is not generally appropriate for factors other than risk factors for the outcome), then the estimated OR will be correct but less precise (i.e., have a wider confidence interval).

A hypothetical case-control study:

Suppose you are conducting an incident case-control study of endometrial cancer and exogenous estrogen. You arrange to be notified of any endometrial cancers diagnosed in 50-70-year-old female, permanent, full-time (or retired and on pension) state employees and retirees in a multi-state area. Assume that all receive medical care benefits; 100,000 are enrolled in fee-for-service plans, and 50,000 are enrolled in managed care (and no one changes!). This population is the study base.

During the five years of follow-up, 200 cases of endometrial cancer develop, for an overall cumulative incidence of endometrial cancer of 133 per 100,000 (0.00133). Of the 200 cases, 175 were exposed to estrogen, and 25 were not (these numbers were derived assuming a cumulative incidence of 200 per 100,000 (0.002) in women with estrogen exposure and 40 per 100,000 (0.0004) in women without exposure, but of course if you knew these incidences, you would not be conducting the study).

Suppose that a much larger percentage (75%) of women in fee-for-service plans are taking exogenous estrogen than are women in managed care (25%). However, you do not know that either, because the prescription records in the various organizations you are dealing with are not computerized (which is why you have resorted to a case-control study rather than following all 150,000 women as a cohort).

For your controls, you first choose a simple random (and by good fortune, precisely representative) sample of 600 women from the 150,000 master file of state employees and retirees. Your data then look as follows:

Southern endometrial cancer and estrogen study (SECES)

	Estrogen	No estrogen	Total
Endometrial cancer	175	25	200
Controls	350	250	600
Total	525	275	800

OR = 5.0
95% confidence interval: (3.19, 7.84)*

*(see chapter on Data Analysis and Interpretation)

(So far so good, since the CIR, by assumption, was 0.002/0.0004 = 5.0.)

However, you are concerned, since you anticipate that estrogen prescribing is very different in the two different types of health care plans. Your suspicion is further supported by the fact that 160/200=80% of the cases are in fee for service, compared to only two-thirds of the random sample controls (400/600) (and 100,000/150,000 in the study base). So even though you have no basis for believing that a woman's health care plan affects her risk for detecting endometrial cancer, you decide to make your control group more like the case group in regard to health plan membership (i.e., you overmatch).

Since 80% of the cases are in fee-for-service and 20% are in managed care, you use stratified random sampling to achieve that distribution in the controls. For 600 controls, that means 480 (80% of 600) from fee-for-service and 120 (20% of 600) from managed care. Since (unbeknownst to you), 75% of the women in fee-for-service take estrogen, as do 25% of the women in managed care, your control group will contain 390 women taking estrogen – 360 exposed women (75% × 480) from fee-for-service and 30 exposed women (25% × 120) in managed care. Thus, your data will now be:

Southern endometrial cancer and estrogen study (SECES)
MATCHED control group

	Estrogen	No Estrogen	Total
Endometrial cancer	175	25	200
Controls	390	210	600
Total	565	235	800

OR = 3.8
95% confidence interval: (2.40, 5.92)

The odds ratio for this table is 3.8, so your matched control group has indeed produced selection bias. Luckily your friend comes by and reminds you that when you use a matched control group, you need to control for the matching factor in your analysis. So you act as if you had conducted two separate studies, one among the women in fee-for-service and the other among the women in managed care (this is called a "stratified analysis" and will be discussed in the chapter on Multicausality – Analysis Approaches). Your two tables (and a combined total for cases and controls) are:

Southern endometrial cancer and estrogen study (SECES)
MATCHED control group, STRATIFIED analysis

	Fee for service			Managed care			Both
	Estrogen	No Estrogen	Total	Estrogen	No estrogen	Total	Grand total
Cancer cases	150	10	160	25	15	40	200
Controls	360	120	480	30	90	120	600
Total	510	130	640	55	105	160	800
OR	5.00			5.00			
95% CI:	(2.55, 9.30)			(2.33, 10.71)			

Stratified analysis* (over both tables): OR=5.0 95% CI: (3.02, 8.73)

* See chapter on Multivariable analysis.

Each of the two case-control studies now has OR = 5.0. The control group within each type of health care plan was a simple random sample. The selection bias in the matched control group held only for the group as a whole compared to the study base as a whole. However, the interval estimate of the OR for the stratified analysis (the last table) is wider than the confidence interval for the OR in the unmatched analysis, indicating a less precise estimate.

Selection bias in cohort studies

Selection bias is generally regarded as a greater danger in case-control than in cohort studies. The reason is that in cohort studies the investigator generally knows how many and which participants were lost to follow-up, so that s/he can assess the potential extent of bias. The investigator can also often examine baseline characteristics of participants who are lost to follow-up for indications that attrition is uniformly distributed and therefore less likely to result in selection bias.

Population cohort attrition

There is, however, a type of attrition that affects both cohort and case-control studies but which is unseen and difficult to categorize. The problem relates to the representativeness of people eligible for study. For simplicity we explain the situation in relation to cohort studies, but since a case-control study is simply an efficient method for studying the same phenomena as a cohort study of the study base, the problem is effectively the same in case-control studies.

A cohort consists of people who are alive at a point or period in calendar time or in relation to some event they undergo (e.g., graduating from college, joining a workforce, undergoing a surgical procedure) and then followed forward in time. The investigators' attention is for the most part directed at what happens after the cohort has been formed, but it is conceivable that mortality and migration occurring before that point have influenced who is available to enroll and thereby influenced what will be observed. If these early selection factors are related to an exposure under study they may diminish an observed effect.

Some examples:

> If a cohort of HIV-infected persons is recruited by enrolling HIV seropositive persons identified through a serosurvey, those who have progressed to AIDS more quickly will be underrepresented as will persons involved in risk behaviors (e.g., injection drug use) that are associated with high mortality. Progression to AIDS in such a cohort will appear different than what would be observed if people were recruited at the time of initial HIV infection.
>
> A study of the effect of hypertension in a cohort of elderly participants cannot enroll persons whose hypertension caused their death prior to the entrance age of the cohort. If those who died earlier had characteristics that made them more vulnerable to end-organ damage from hypertension, then the cohort study may observe less morbidity and mortality associated with hypertension than would be observed if the study had enrolled younger participants.
>
> Even a cohort study in a population of newborns can enroll only infants from conceptions that result in a live birth. If environmental tobacco smoke (ETS) increases the rate of early fetal losses, possibly undetected, there may be differences between the fetuses who die and those who survive to birth. If fetuses who survive are more resistant to harm from ETS, then a cohort study of harmful effects of ETS on infants may observe a weaker effect because the most susceptible of the exposed cases were never enrolled in the cohort.

Assuming that the target populations are defined as persons age 70 years and older (in the hypertension study) or newborns (in the ETS study), then internal validity as defined above would not appear to be affected. But study findings could nevertheless be misleading. If cholesterol lowering medication lengthens disease-free life in hypertensives, then more hypertensives taking this medication will survive to age 70 to enter the cohort. If these hypertensives have a higher rate of developing end-organ damage, then the observed rate of morbidity and mortality associated with

hypertension will be higher, people taking cholesterol-lowering medication may now be observed to have higher morbidity and mortality, and the stronger effect of hypertension will be found to be associated with cholesterol-lowering medication. Similarly, a factor that reduces fetal loss in ETS-exposed pregnancies will increase the proportion of ETS-susceptible infants enrolled in the cohort study and will be associated with higher infant morbidity/mortality.

This problem would appear to be closer to lack of external validity (generalizability across time or setting), but it bears a strong resemblence to selective survival as encountered in a cross-sectional or case-control study (e.g., the example of a case-control study of aggressiveness and MI, used above). Thus losses prior to the inception of a cohort need careful consideration so that the investigator is not misled by selective factors operating at an earlier stage.

Selection bias due to missing data

One other potential cause of selection bias in studies of all kinds is missing data for a variable required in the analysis. Bias due to missing data is usually a topic considered under the heading of analysis, but its effect is akin to selection bias and its prevention requires avoidance of systematic differences in rates of missing data.

The problem can be particularly severe in analyses involving a large number of variables. For example, regression procedures often exclude an entire observation if it is missing a value for any one of the variables in the regression. This practice (called "listwise deletion") can exclude large percentages of observations and induce selection bias, even when only 5% or 10% of missing values for any one variable. Imputation procedures can often avoid the exclusion of observations, and depending upon the processes that led to the missing data (the missing data "mechanism") they can lead to less or unbiased analyses. There are also analytic procedures that can reduce the bias from nonresponse (inability to enroll participants) and/or attrition (loss of participants following enrollment).

Information bias

Information bias refers to systematic distortion of estimates resulting from inaccuracy in measurement or classification of study variables (misclassification bias is a subcategory of information bias when the variable has only a small number of possible values). For example, a disease may be present but go unrecognized, a blood pressure may be misread or misrecorded, recall of previous exposure may be faulty, or in extreme cases, data may have simply been fabricated by uncooperative subjects or research personnel. Typical sources of information/misclassification bias are:

1. variation among observers and among instruments – or variation across times by the same observer or instrument;
2. variation in the underlying characteristic (e.g, blood pressure) – and that variation has not been adequately accomodated by study methods;
3. misunderstanding of questions by a subject being interviewed or completing a questionnaire – or inability or unwillingness to give the correct response; or selective recall;
4. incomplete or inaccurate record data.

Systematic overview:

Information bias can occur with respect to the disease, the exposure, or other relevant variables. Sometimes, information bias can be measured, as when two methods of measurement are available, one being deemed more accurate than the other. Sometimes, information bias can be assumed to exist but cannot be directly assessed.

For example, if there is a true causal relationship between estrogens and endometrial cancer, i.e., a biological process by which estrogen molecules initiate or promote cancerous cell growth in the endometrium, then this pathophysiologic process presumably relates to certain specific molecular species, operating over a certain time period, and resulting in certain forms of endometrial cancer. To the extent that endometrial cancer is a heterogeneous entity, and the estrogen-related form is one subtype, then the observed association between endometrial cancer and estrogen is being diluted, as it were, by combining in one case group cancers caused by estrogens and cancers resulting from other mechanisms. Masking of the relationship also occurs by combining in one exposure group women whose exposure caused their cancers, women whose exposure to estrogen occurred only before or after the relevant time period in terms of the natural history of endometrial cancer, and women who were exposed to a nonpathogenic form of estrogen, nonpathogenic dose, or nonpathogenic mode of administration (should there be such).

Another example is the study of the health effects of exposure to lead. The traditional index of absorption, blood lead level, reflects only recent exposure, because the half-life of lead in blood is only about 36 days (see Landrigan, 1994). So there may be relatively little relationship between a single blood lead measurement and body lead burden. Pioneering studies by Herbert Needleman

employing lead exposure measures from deciduous teeth enabled the demonstration of a relationship between low lead exposure and cognitive and behavioral impairment in children. Now, the advent of K x-ray florescence analysis of lead in bone, where the half life is on the order of 25 years, may provide an important new tool in epidemiologic studies of lead exposure (Kosnett MJ et al., 1994).

For these reasons, rigorous study design and execution employ:

1. verification of case diagnosis, employing such procedures as multiple independent review of tissue samples, x-rays, and other diagnostic data;
2. definition of homogeneous subgroups, with separate analysis of data from each;
3. multiple data sources concerning exposure (and other relevant variables), permitting each to corroborate the other;
4. precise characterization of actual exposure, with respect to type, time period, dosage, etc.

Unfortunately, reality constraints impose compromises. For example, data from 20 years ago may be the most relevant in terms of the causal model, but data from two years ago may be much more available and accurate. In using the more recent data, one either assumes that recent exposure is a good proxy measure for previous exposure or that the recent exposure is also related to the disease, though perhaps not as strongly as the previous exposure.

[For more on the above, see Hulka, Hogue, and Greenberg, "Methodologic issues in epidemiologic studies of endometrial cancer and exogenous estrogen", and Kenneth J. Rothman, Induction and latent periods, *Am J Epidemiol* 114:253-259, 1981 (the Rothman article addresses the question of timing of exposure).]

One consideration raised by the above is the importance of developing specific hypotheses in advance of the study. Such hypotheses, if they can be elaborated, strengthen both the design and interpretation of the study. The design is strengthened because the hypotheses guide the investigator in selecting the relevant variables and their features (time of occurrence, etc.) on which to obtain data. It may not be possible to obtain just the right information, but at least the hypotheses protect guide the search. Hypotheses also provide guidance about what relationships to analyze and how to construct analysis variables (e.g., what disease subcategories to relate to which forms of exposure). Specific hypotheses – grounded in existing knowledge and theory – can also increase the persuasiveness of the findings.

Basic terms and concepts

Reliability (of measurement or classification) concerns the repeatability of a measurement – across time, across measurement instruments, across observers. If a measure is reliable, it may still not be accurate. But if a measure is not reliable, then the data values for it contain a substantial random component. This random component reduces the information content of the variable, the strength of associations involving it, and its effectiveness in controlling confounding (to be discussed in a

following chapter). The concept of reliability is relevant when two or more measures of comparable authoritativeness are being compared.

Validity (of measurement or classification) is the extent to which a measurement measures what it is supposed to measure. Assessment of validity, therefore, implies the availability of a measurement method that can be regarded as authoritative (often called a "gold standard"). Since one measure has more authority, our interest shifts from simple agreement between measures to evaluation of the less authoritative one. For example, if mean blood pressure measured with a random-zero mercury sphygmomanometer over a series of readings in a person lying on his/her back is our standard for "true" blood pressure, then a casual pressure in a person sitting down will be systematically inaccurate, since it will tend to be higher. Although we will examine agreement between the supine and casual blood pressures, our interest is on the accuracy of the latter with respect to the "gold standard".

Relationship of reliability and validity

"Validity" is used as a general term for accuracy or correctness. The procedure of assessing the accuracy of a measurement instrument is often referred to as validation. In many situations, though, we do not know the correct result, so the best we can do is to compare measurements that are assumed to be equally accurate. In these situations, agreement between measurements is termed "reliability".

In this sense, reliability is a subcategory of validity. However, reliability (repeatability, consistency) can be present without validity (two faculty can agree completely yet both be wrong!). Also, a measurement procedure can be valid in the sense that it gives the correct value on average, though each measurement includes a large amount of random variation (e.g., 24-hour dietary recall for cholesterol intake). Sometimes it is said that "a measure that is unreliable cannot be valid". Whether this statement is true or not depends upon what aspect of validity is being considered. More commonly, random error (unreliability) and bias (lack of validity) are regarded as independent components of total error.

Psychometrics is the subdiscipline of psychology that addresses the evaluation of questionnaire items and scales. "Validation" as used in psychometrics encompasses both reliability (consistency) and validity. However, due to the scarcity of classifications and measures that can be regarded as authoritative, much of psychometric validation concerns assessment of reliability. Common situations where reliability is important to examine are comparisons of performance of two raters or interviewers of equal stature (inter-rater reliability), of results from repeated measurements of a characteristic that is believed to be stable (test-retest reliability), and of scores from equivalent items that make up a scale (inter-item reliability – often termed "internal consistency").

Assessment of reliability

Validation involves the measurement of agreement between two or more measurements or classifications. Agreement is not identical to "association", but is rather a special case – the case in

which the both measures increase in the same direction and have the same scale. (An obvious example of an association which does not imply agreement is an inverse association.)

Percent agreement

For categorical variables, a simple measure of reliability is the percentage of instances in which the two measurement instruments agree. Supposed that 100 electrocardiograms (ECG) are given to two expert readers to code independently as "abnormal" or "normal", and that the two readers agree on 90 (30 that they both call "abnormal" and 60 that they both call "normal"). But the 90 percent agreement is not as good as it seems, since it gives "credit" for agreement that we would expect to occur just by chance. What if the two readers, preferring to play golf, left the ECG's with their secretaries, with instructions for each secretary to code each ECG independently by rolling a die and coding the ECG as "abnormal" if the die came up 6. To the unfortunate investigator, when s/he checked the reliability of the coding, the two readers might appear to have 72% agreement (3 abnormals and 69 normals).

Categorical variables - Kappa

One measure of reliability that adjusts for agreement expected to occur by chance is ***Kappa*** (introduced by Cohen in 1960). For a categorical variable without inherent ordering (e.g., initial diagnosis of chest pain as ?angina, ?gastroesopheal reflux, or ?musculoskeletal), Kappa is computed as:

$$K = \frac{p_o - p_c}{1 - p_c}$$

p_o = observed proportion of agreement

p_c = proportion of agreement expected by chance

The proportion of agreement expected by chance is computed by using the marginal percentages, the same procedure that is used for computing a chi-square test for associatiion.

Suppose that a managed care organization (MCO) is investigating the reliability of physician diagnostic work-up for chest pain. Two physician's initial assessments and order for diagnostic testings are compared for 100 sequential patients presenting with uncomplicated occasional chest pain at their initial visit to the MCO.

Comparison of diagnoses by physicians A and B among 100 patients reporting chest pain

		Physician B			
Physician A		?Angina	?Reflux	?Musculo-skeletal	Total
	?Angina	12	1	1	14
	?Reflux	2	36	4	42
	?Musculo-skeletal	2	8	34	44
	Total	16	45	39	100

Since the physicians agree on the initial diagnosis for 12 + 36 + 34 patients, their percent agreement is 82/100 = 82%. However, based on the marginals we expect considerable agreement by chance alone. The expected proportion of agreement by chance is computed from the marginal distributions as follows:

Expected proportion of agreement =

	(Proportion ?Angina Physician A) × (Proportion ?Angina Physician B)	14/100 × 16/100
+	(Proportion ?Reflux Physician A) × (Proportion ?Reflux Physician B)	42/100 × 45/100
+	(Proportion ?Mus-Sk Physician A) × (Proportion ?Mus-Sk Physician B)	44/100 × 39/100
=	14/100 × 16/100 + 42/100 × 45/100 + 44/100 × 39/100	
=	0.0224 + 0.189 + 0.1716 = 0.383	

The value of Kappa for this table is therefore: $$K = \frac{0.82 - 0.383}{1 - 0.383}$$

For assessing agreement between ordinal variables with few categories, weighted versions of Kappa are used in order to assign varying weights to different degrees of disagreement. A discussion of Kappa may be found in Joseph Fleiss's text *Statistical Methods for Rates and Proportions*. The second edition suggests adjectives for characterizing values of Kappa.

Continuous variables

For continuous measures and ordinal variables with many categories, the data display is a scatterplot, rather than a crosstabulation. Perfect agreement means that all of the measurement pairs lie on a straight line with slope 1 and intercept 0 (i.e., the line goes through the origin). The most direct index of the level of agreement between the two measures are the regression coefficient and intercept for the straight line that best fits the measurement pairs. The closer the regression

coefficient (slope) is to 1.0 and the regression intercept is to zero, and the narrower their confidence intervals, the better the level of agreement.

A common index of agreement is the correlation coefficient. The product-moment (Pearson) correlation coefficient (r) assesses the extent to which pairs of observations from the two measurements lie on a straight line. The Spearman (rank) correlation coefficient, *rho*, used for ordinal variables, assesses the extent to which the pairs of observations have the same ordering for the two measurement instruments (the lowest for the first instrument is close to the bottom for the second instrument, the tenth lowest for the the first is close to being tenth lowest for the second, and so on).

However, correlation coefficients ignore location and scaling. Thus, if the readings from one thermometer are always exactly two degrees below the readings from a second thermometer, agreement is certainly less than imperfect, yet the correlation coefficient between their readings will be 1.0 (for a perfectly straight line of slope 1.0, but not through the origin). If the readings from the first thermometer are always twice those of the second, the correlation will also be 1.0 (for a straight line through the origin, but with a slope of 2.0). Therefore a correlation coefficient alone is an inadequate assessment of agreement. It must be accompanied by a comparison of the location (mean, median) and scale (standard deviation or other measure of dispersion) for the readings of the two measures.

Reliability of a scale [optional]

A measure of reliability that is widely used in psychometrics is Cronbach's coefficient alpha. Coefficient alpha's conceptual basis (see Nunnally, *Psychometric Theory*) can be stated as follows.

Suppose that you have a set of questionnaire items each of which attempts to measure the same, unobservable construct (a "latent variable"). The response value for any individual item will reflect the value of that latent variable but also some amount of error, which is assumed to be random, independent of everything else, and symmetrically distributed with mean zero. Under these assumptions, the average of the response values for the set of items will provide a more reliable measure of the latent variable than is available from any single item (just as the average value for a set of otherwise equivalent blood pressure measurements will yield a more accurate (precise) value than any single measurement). The random components in the item responses should counterbalance each other, so that the average is a more precise measure of the latent variable.

In such a scenario, coefficient alpha assesses how much of the scale scores reflect the values of the latent variable and how much reflects measurement error. The the higher the "shared item variance" (the more the individual items in the scale agree with each other) and the larger the number of items, the higher the value of alpha. More precisely stated, coefficient alpha is the proportion of the total variance in the scale scores that represents the variance of the values of the latent variable (the rest being the variance of the random errors for each

item). Alpha values of 0.80 are are considered adequate for computing correlations and fitting regression models, and a sample size of 400 observations is regarded as adequate to estimate alpha (see Nunally).

Obstacles to realizing this ideal scenario include the probability that items are not perfectly equivalent, that people's responses to some items in the scale affect their answers to other items (so errors in item responses are not independent), and that factors other than the latent variable contribute non-random variation in item responses (thereby introducing systematic error, i.e., bias). Note that coefficient alpha does not address bias, only random variability.

Assessment of validity – sensitivity and specificity

As noted above, assessment of validity is directed toward the evaluation of a rater or measurement instrument compared to an authoritative rater or instrument. For detection of a characteristic or a condition, epidemiologists generally employ the concepts of sensitivity and specificity that were introduced in a previous chapter. Using the words "case" ("noncase") to refer, respectively, to people who have (do not have) the condition or characteristic (e.g., a disease, an exposure, a gene) being measured, then sensitivity and specificity are, respectively, the probabilities for correctly classifying cases and noncases.

Sensitivity is the ability to detect a case.

Specificity is the ability to detect a noncase.

Example:

If a procedure correctly identifies 81 of 90 persons with a disease, condition, or characteristic, then the sensitivity of the procedure is:

$$Se = 81/90 = 0.9 = 90\%$$

If the same procedure correctly identifies 70 of 80 persons without the disease, condition, or characteristic, then the specificity of the procedure is:

$$Sp = 70/80 = 0.875 = 88\%$$

In probability notation,

$$Se = Pr(D' \mid D) \qquad Sp = PR(\bar{D}' \mid \bar{D})$$

where D = case, $\bar{D}$ = noncase, D' = "classified as a 'case'", and $\bar{D}'$ = "classified as a 'noncase'".

The inverse of sensitivity and specificity are "false negatives" and "false positives". Some authors prefer to avoid the latter terms, because of the potential confusion about whether "negative" and "positive" refer to the test (in accordance with the definition in John Last's *Dictionary of Epidemiology* or to the true condition. However, the terms remain in common use, and we will follow the Dictionary's usage, whereby a "false negative" is a negative test result in a person who has the characteristic (i.e., an erroneous negative test) and "false positive" is an erroneous positive test result.

Sensitivity and specificity as defined above suffer from the same limitation that we have noted for percent agreement, that their calculation fails to take account of agreement expected on the basis of chance. Even a random process will classify some cases and noncases correctly. Methods for dealing with this limitation have been published (Roger Marshall, "Misclassification of exposure in case-control studies", Epidemiology 1994;5:309-314), but are not yet in wide use.

Impact of misclassification

The impact of misclassification on estimates of rates, proportions, and measures of effect depend on the circumstances. Consider the following example for a rare disease. Assume a cohort of 1,000 participants, of whom 60 develop CHD during the a four-year interval.

If the sensitivity of our diagnostic methods is only 0.80 (or 80%), then we will detect only 48 of those cases (48/60, i.e., 80% of 60). There will be 12 false negatives.

If the specificity of our diagnostic methods is 0.90 (or 90%), then we will incorrectly classify 94 of the 940 subjects who did not develop CHD (90% of the 940 noncases will be correctly identified as such, leaving 94 (940 minus 846) noncases to be incorrectly classified as "cases"). These 94 subjects will be false positives.

Thus, we will observe (or think we observe) 142 "cases" of CHD – 48 who in fact have CHD and 94 who actually do not. Note that in this case the majority of "cases" do not have the disease! This example illustrates the dilemma of false positives when studying a rare disease. The false positives and their characteristics will "dilute" or distort the characteristics of any "case" group we might assemble. Hence the emphasis on avoiding false positives through case verification, using such methods as pathological confirmation.

Suppose that the participants in this cohort are "exposed", and another similar cohort consists of 1,000 participants who are not "exposed". Assuming that the diagnostic accuracy is not influenced by exposure status, we expect the results for the two cohorts to be as follows:

Hypothetical scenario showing effect of misclassification bias on measures of association

	True (Se=1.0, Sp=1.0)		Observed: Se=0.8, Sp=1.0		Observed: Se=1.0, Sp=0.9		Observed: Se=0.8, Sp=0.9	
	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$	E	$\bar{E}$
D	60	30	48	24	154	127	142	121
$\bar{D}$	940	970	952	976	846	873	858	879
	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
RR	2.0		2.0		1.21		1.17	
RD	0.03		0.024		0.027		0.021	

From this example, we can see that:

1. Even rather high levels of sensitivity and specificity do not avoid bias;
2. Different epidemiologic measures are affected in different ways;
3. The RR need not be affected by imperfect sensitivity if specificity is sufficiently high; the RD will be affected, though.
4. The RR will be affected by imperfect specificity for detecting a rare disease even if sensitivity is high; however, the RD may be affected only slightly.
5. Specificity is of utmost importance for studying a rare disease, since it is easy to have more false positive tests than real cases, identified or not;
6. Bias in the classification of a dichotomous disease typically masks a true association, if the misclassification is the same for exposed and unexposed groups.

[Try creating a spreadsheet to see how various levels of sensitivity and specificity change the RR and RD. Convenient formulae are in the appendix.]

Types of misclassification

The above example deals with misclassification of a disease, misclassification that is independent of exposure status. Nondifferential misclassification of a dichotomous exposure variable, i.e., misclassification that occurs independently of disease status – will bias ratio measures of effect toward the null value of 1.0. This will also be the case for nondifferential misclassification of a dichotomous disease variable or of both a dichotomous disease variable and a dichotomous exposure simultaneously.

Differential misclassification, however, where errors in measuring one variable vary according to the value of another variable, can lead to bias in any direction. Common scenarios for differential misclassification are selective recall of past exposures or selective detection of disease based on knowledge of the patient's exposure history. Also, when the misclassified variable has more than two levels, even nondifferential misclassification can produce bias in any direction (because this last point has been emphasized only in recent years and because traditionally the teaching of epidemiology has focused on dichotomous disease and exposure variables, it is not uncommon to hear the maxim "nondifferential misclassification bias is towards the null" without mention of the exceptions).

In addition, measurement error for other variables involved in the analysis produces bias in a direction that depends on the relationships of the variables. For example, if we are performing age adjustment and have bias in the measurement of age, then the age adjustment will not completely remove the effect of age. A situation of this type is referred to as information bias in the measurement of a covariable and is discussed in Rothman and Greenland.

Direction and extent of bias

The importance of being able to discern the direction of the bias and, if possible, to assess its magnitude, is to enable interpretation of the observed data. For example, if a positive association is observed between two factors and the direction of misclassification bias can be shown to be toward the null, then such bias could not be responsible for the finding of a positive association. Similarly, if misclassification bias can be shown to be in the positive direction, then the failure to find an association cannot be due to that bias. In addition, techniques exist to correct for errors in measurement in a number of analytic procedures. However, these procedures often require some outside estimate of sensitivity and specificity.

Where the categories of bias break down

Earlier it was mentioned that the boundaries between random error and systematic error as well as those among the three classes of bias sometimes become unclear. Here are some situations that are challenging to classify.

False negatives in detecting cases for a case-control study.

If cases are missed due to information bias, then such persons will not be counted as cases in a case-control study. If this lack of sensitivity is in some way related to exposure status (e.g., greater detection of endometrial cancer among women who take estrogen – the detection bias issue), then the case group will not be representative of the population of cases.

From the viewpoint of the case-control study this type of bias will be classified as selection bias, since it is manifest through differential selection probabilities for cases. But the error mechanism in this scenario was misclassification of cases. Moreover, if some women with asymptomatic endometrial cancer happen to be selected as controls, their presence in the control group is

presumably classified as information (misclassification) bias, since in this situation the subjects belong in the study, except that they should be in the case group.

Variability in a parameter being measured can produce both random error (measurement imprecision) and information bias in a measure of effect

Blood pressure, for example, varies from moment to moment, so that every measurement of blood pressure reflects a degree of random variability (random error). If blood pressures are measured on a single occasion, and then disease incidence or some other endpoint is recorded during the ensuring five years, the observed association between blood pressure and the outcome will understate any true association.

The reason for this is that subjects who were classified as "high" on their initial measurement will include some who were "high" just by chance. Subjects classified as "low" will include some who were "low" just by chance. The resulting error in exposure measurement will muddy the contrast between the group outcomes compared to what would obtain if our "high" group contained only those who were truly "high" and low group contained only those who were truly "low".

Assuming that chance variability is independent of study outcomes, then the result is nondifferential misclassification, and the observed association will be weaker than the "true" association. Thus, random error can produce systematic error, or bias.

Blood pressure (and other physiological parameters) also varies in a diurnal pattern, being lower in the morning and rising during the day. Failure to provide for diurnal variation can produce several kinds of error. For example, if blood pressures are measured on study subjects at random times during the day (i.e., each subject's blood pressure is measured once, but any given subject may be examined at any time of day), then the diurnal variation adds a component of random error to that from the moment to moment variation. Therefore, estimates of group means and their differences will be more imprecise than if measurements had been conducted at the same time of day.

If for some reason subjects in one category (e.g., blue collar workers) are examined in the morning and subjects in another category (e.g., homemakers) are examined in the afternoon, then there will be a systematic difference between the mean blood pressures for subjects in the different categories, a systematic difference arising from the systematic variation in blood pressure combined with the systematic difference in time of measurement. The resulting systematic error could lead to selection bias or information bias depending upon the nature of the study.

Regression to the mean

A well-known phenomenon that illustrates how random variability can lead to systematic error is regression to the mean (Davis CE: The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol* 104:493-498, 1976). When a continuous variable, such as blood pressure or serum cholesterol, has a degree of random variability associated with it (or with its measurement), then each measurement can be thought of as based on the "true value" for the

subject plus or minus a random noise factor. If the distribution of the random variable is symmetric with a mean of zero, then the average value of a series of different readings will be close to the "true value". If the random component is large, however, any given measurement can be substantially above or below the average.

In such a situation, a variable for which a given measurement falls at the high end or the low end of the distribution for that variable will tend to be closer to the center of the distribution for a later measurement. For example, in the Lipid Research Clinics (LRC) Prevalence Study, populations were screened for cholesterol and triglyceride levels, and those with elevated levels were asked to return for additional evaluation. If, say, 15% of the subjects screened were asked to return, it can be expected (and did happen) that many of those subjects did not have elevated levels upon re-measurement.

The reason for this "regression" is that the group of subjects in the top 15% of the lipid distribution at their screening visit consists of subjects whose lipid measurement was high due to a large positive random component as well as subjects whose lipid levels are truly high. On re-measurement, the random component will, on average, be smaller or negative, so that subjects without truly high lipid levels will fall below the cutpoint as well as some subjects with truly high levels but who on this measurement have a large negative random component.

If by an extreme value we mean one that is "unusually high", that implies that usually it should be lower. The opposite is true for unusually low values. Therefore, the average serum cholesterol in an unselected population will not tend to "regress towards the mean", since in a random process the increases and decreases will balance each other. But if we select a portion of the population based on their initial measurements' being high (and/or low), then that selected population will tend to "regress" towards the population mean.

In regression toward the mean, we have a situation in which random variability can produce systematic distortion, in the sense that the mean of the cholesterol levels (or blood pressures) of the "elevated" subjects overstates their "true mean" (assuming that "true" is defined as an average of several measurements). Whether this distortion produces selection bias or information bias will depend upon the actual process of the study.

Suppose that "high risk" subjects (elevated cholesterol, blood pressure, and other CVD risk factors) are enrolled in a "wellness" program and their risk levels are measured several months later, there will probably be some decline in these levels regardless of the program's effects, simply due to regression to the mean. This process is one reason for the importance of a randomly allocated control group, which would be expected to experience the same regression.

[According to John R. Nesselroade, Stephen M. Stigler, and Paul B. Baltes, regression to the mean is not a ubiquitous phenomenon, but depends upon the characteristics of the underlying model or process involved. A thorough, but largely statistical, treatment of the topic can be found in "Regression toward the mean and the study of change," Psychological Bulletin 88(3):622-637, 1980.]

Appendix 1

Formulas to see the effects of various levels of sensitivity and specificity on the RR and RD

If a, b, c, d are the TRUE values of the cells in a four-fold table, then the observed RR and observed RD in the presence of Sensitivity (Se) and Specificity (Sp) for measuring disease are given by:

$$\text{Observed RR} = \frac{[(Se)a + (1\text{-}Sp)c]/n_1}{[(Se)b + (1\text{-}Sp)d]/n_0}$$

$$\text{Observed RD} = \frac{(Se)a + (1\text{-}Sp)c}{n_1} - \frac{(Se)b + (1\text{-}Sp)d}{n_0}$$

$$= Se\left(\frac{a}{n_1} - \frac{b}{n_0}\right) + (1 - Sp)\left(\frac{c}{n_1} - \frac{d}{n_0}\right)$$

Appendix 2

More on the concern to avoid false positive diagnoses of disease in case-control studies of a rare disease (e.g., endometrial cancer and estrogen) – the importance of verification of case status: [This is a simplified version of the presentation in the Hulka, Hogue, and Greenberg article in the bibliography.]

The case-control strategy aims to estimate the probability of exposure in cases and in noncases, the ideal for the latter being the general population from which the cases arose. Misclassification of disease leads to contamination of these probability estimates. In particular, false positives "dilute" the cases:

The observed probability of exposure in subjects classified as "cases" equals:

1. the probability of exposure in true cases
2. plus a distortion equal to the proportion of false positive "cases" multiplied by the difference in exposure probability between true cases and false positives.

Algebraically,

$$\Pr(E|D') = \quad \text{— the } \textbf{observed} \text{ exposure prevalence in "cases"}$$

$$\Pr(E|D) \quad \text{— the } \textbf{true} \text{ exposure prevalence in cases}$$

$$+ \Pr(\bar{D}|D')\,[\Pr(E|\bar{D}) - \Pr(E|D)] \quad \text{— the } \textbf{bias}$$

where E = exposure, D = a **true** case, $\bar{D}$ = a **true** noncase,
and D' = **any** subject who is classified as a "case" (correctly or incorrectly)

So $\Pr(\bar{D}|D')$ is the probability that someone who is **called** a "case" is in fact a **noncase**

and $\Pr(E|\bar{D})$ is the probability that a **true noncase** has the exposure.

Correspondingly, the **observed** probability of exposure in subjects classified as "**noncases**" equals:

1. the probability of exposure in true **noncases**
2. plus a distortion equal to the proportion of false negatives among persons classified as "noncases" multiplied by the difference in exposure probability between true noncases and false negatives.

Algebraically,

$\Pr(E|\bar{D}') =$ — the **observed** exposure prevalence in "noncases"

$\Pr(E|\bar{D})$ — the **true** exposure prevalence in noncases

$+ \Pr(D|\bar{D}')\ [\Pr(E|D) - \Pr(E|\bar{D})]$ — the **bias**

where $\bar{D}'$ = any subject classified as a "noncase" (correctly or incorrectly)

Numerical example:

If:

the probability of exposure in true cases = 0.4,

the probability of exposure in true noncases = 0.2,

the probability of exposure in false positives = 0.2 (i.e., the false positives are really just like other noncases)

then in a sample of subjects classified as "cases" in which one-third are falsely so classified (i.e., false positives) we expect to observe a probability of exposure of:

$$\Pr(E|D') = 0.4 + (1/3)\ [0.2\text{-}0.4] = 0.4 - (1/3)[0.2] = 0.333$$

or equivalently,

$$\Pr(E|D') = (2/3)(0.4) + (1/3)(0.2) = 0.333$$

(i.e., the prevalence of exposure is a weighted average of the exposure prevalence in the correctly classified cases and the exposure prevalence in the false positives).

Since the true probability of exposure in cases is 0.4, the observed results are biased downward. Since the proportion of false negatives in the control group (diseased subjects classified as "noncases") will generally be small if the disease is rare, the estimate of the probability of exposure in noncases will generally not be biased.

The true OR is 2.67 [{.4/(1-.4)}/{.2/(1-.2)}]; the observed OR is 2.0 [{.333/(1-.333)}/{.2/(1-.2)}]. The discrepancy would be greater if the true exposure probabilities were more different.

Bibliography

Henneckens and Buring. *Epidemiology in Medicine*. Rothman and Greenland, 1998. Rothman. *Modern Epidemiology*. Chapters 7, 8. Kleinbaum, Kupper and Morgenstern. *Epidemiologic Research: Principles and Quantitative Methods*. Chapter 10, Introduction to Validity.

Armstrong BK, White E, Saracci Rodolfo. Principles of exposure measurement in epidemiology. NY, Oxford, 1992, 351 pp., $59.95. Key reference (reviewed in Am J Epidemiol, April 15, 1994).

Brenner, Hermann and David A. Savitz. The effects of sensitivity and specificity of case selection on validity, sample size, precision, and power in hospital-based case-control studies. Am J Epidemiol 1990; 132:181-192.

Cohen, Bruce B.; Robet Pokras, M. Sue Meads, William Mark Krushat. How will diagnosis-related groups affect epidemiologic research? Am J Epidemiol 1987; 126:1-9.

Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? Am J Epidemiol 1990; 132:746-8; and correspondence in 1991;134:441-2 and 135(12):1429-1431

Feinleib, Manning. Biases and weak associations. Preventive Medicine 1987; 16:150-164 (from Workshop on Guidelines to the Epidemiology of Weak Associations)

Feinstein AR, Horwitz RI. Double standards, scientific methods, and epidemiologic research. New Engl J Med 1982; 307:1611-1617.

Feinstein, Alvan R.; Stephen D. Walter, Ralph I. Horwitz. An analysis of Berkson's Bias in case-control studies. J Chron Dis 1986; 39:495-504.

Flanders, W. Dana; Harland Austin. Possibility of selection bias in matched case-control studies using friend controls. *Am J Epidemiol* 1986; 124:150-153.

Flanders, W. Dana; Coleen A. Boyle, John R. Boring. Bias associated with differential hospitalization rates in incident case-control studies. *J Clin Epidemiol* 1989; 42:395-402 (deals with Berkson's bias for incident case control studies - not a major work)

Flegal, Katherine M., Cavell Brownie, Jere D. Haas. The effects of exposure misclassification on estimates of relative risk. Am J Epidemiol 1986; 123:736-51.

Flegal, Katherine M.; Penelope M. Keyl, and F. Javier Nieto. Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol 1991; 134(10):1233-44.

Gregorio, David L.; James R. Marshall, Maria Zielezny. Fluctuations in odds ratios due to variance differences in case-control studies. *Am J Epidemiol* 1985; 121:767-74.

Horwitz, Ralph I. Comparison of epidemiologic data from multiple sources. J Chron Dis 1986; 39:889-896.

Horwitz, Ralph I. and Alvan R. Feinstein. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. N Engl J Med 1978;299:1089-1094.

Hulka, Barbara S., Carol J.R. Hogue, and Bernard G. Greenberg. Methodologic issues in epidemiologic studies of endometrial cancer and exogenous estrogen. Amer J Epidemiol 1978; 107:267-276.

Hulka, BS, Grimson RC, Greenberg BG, et al. "Alternative" controls in a case-control study of endometrial cancer and exogenous estrogen. Am J Epidemiol 1980;112:376-387.

Hutchison, George B. and Kenneth J. Rothman. Correcting a bias? Editorial. N Engl J Med 1978;299:1129-1130.

Kosnett JM, Becker CE, Osterich JD, Kelly TJ, Pasta DJ. Factors influencing bone lead concentration in a suburban community assessed by noninvasive K X-ray fluorescence. JAMA 1994;271:197-203

Landrigan, Philip J. Direct measurement of lead in bone: a promising biomarker. Editorial. JAMA 1994;271:239-240.

Maclure, Malcolm; Walter C. Willett. Misinterpretation and misuse of the Kappa statistic. *Am J Epidemiol* 1987; 126:161-169.

Marshall, James R. and Saxon Graham. Use of dual responses to increase validity of case-control studies. *J Chron Dis* 1984;37:107-114. (also commentary by Stephen D. Walter, authors' reply, and Walter's reply in that issue).

Neugebauer, Richard and Stephen Ng. Differential recall as a source of bias in epidemiologic research. *J Clin Epidemiol* 1990; 43(12):1337-41.

Nunnally, Jum C. and Ira H. Bernstein. *Psychometric theory*. New York: McGraw-Hill, 1994.

Roberts, Robin S., Walter O. Spitzer, Terry Delmore, David L. Sackett. An empirical demonstration of Berkson's bias. *J Chron Dis* 1978; 31:119-128.

Sackett, D.L.: Bias in analytic research. *J Chron Dis* 32:51-63, 1979. (and comment). In Ibrahim: *The Case-Control Study*.

Schatzkin, Arthur; Eric Slud. Competing risks bias arising from an omitted risk factor. *Am J Epidemiol* 1989; 129:850-6.

Sosenko, Jay M.; Laurence B. Gardner. Attribute frequency and misclassification bias. *J Chron Dis* 1987; 40:203-207.

Walker, Alexander M. Comparing imperfect measures of exposure. *Am J Epidemiol* 1985; 121:783-79

Warnecke RB, Johnson TP, Chavez N, Sudman S, O'Rourke DP, Lacey L, Horm J. Improving question wording in surveys of culturally diverse populations *Ann Epidemiol* 1997; 7:334-342.

Warnecke RB, Sudman S, Johnson TP, O'Rourke D, Davis AM, Jobe JB. Cognitive aspects of recalling and reporting health-related events: Pap smears, clinical breast exams, and mammograms. *Am J Epidemiol* 1997; 13(3):305-315.

White, Emily. The effect of misclassification of disease status in follow-up studies: implications for selecting disease classification criteria. Am J Epidemiol 1986; 124:816-825.