UNIT 2.

BIOMARKERS: PRACTICAL ASPECTS

CHAPTER 8.

Measurement error in biomarkers: Sources, assessment, and impact on studies*

Emily White

*Parts of this chapter appear in White E. Effects of biomarker measurement error on epidemiological studies. In: Toniolo P, Boffetta P, Shuker DEG, Rothman N, Hulka B, Pearce N, editors. *Applications of biomarkers in cancer epidemiology.* Lyon, IARC Scientific Publication; 1997. p. 73–94

Summary

Measurement error in a biomarker refers to the error of a biomarker measure applied in a specific way to a specific population, versus the true (etiologic) exposure. In epidemiologic studies, this error includes not only laboratory error, but also errors (variations) introduced during specimen collection and storage, and due to day-to-day, month-to-month, and year-to-year within-subject variability of the biomarker. Validity and reliability studies that aim to assess the degree of biomarker error for use of a specific biomarker in epidemiologic studies must be properly designed to measure *all* of these sources of error. Validity studies compare the biomarker to be used in an epidemiologic study to a perfect measure in a group of subjects. The parameters used to quantify the error in a binary marker are sensitivity and specificity. For continuous biomarkers, the parameters used are bias (the mean difference between the biomarker and the true exposure) and the validity coefficient (correlation of the biomarker with the true exposure). Often a perfect measure of the exposure is not available, so reliability (repeatability) studies are conducted. These are analysed using kappa for binary biomarkers and the intraclass correlation coefficient for continuous biomarkers. Equations are given which use these parameters from validity or reliability studies to estimate the impact of nondifferential biomarker measurement error on the risk ratio in an epidemiologic study that will use the biomarker. Under nondifferential error, the attenuation of the risk ratio is towards the null and is often quite substantial, even for reasonably accurate biomarker measures. Differential biomarker error between cases and controls can bias the risk ratio in any direction and completely invalidate an epidemiologic study.

Introduction

Importance of understanding the degree of measurement error in biomarkers

When a biomarker is being considered for use in an epidemiologic study, or has been selected, the researcher needs to become familiar with its measurement properties (i.e. how well the measure selected reflects the underlying exposure of interest). There are many sources of error in biomarkers when they are used in epidemiologic studies. These include not only laboratory error, but also errors due to variation in the specimen collection and processing methods, as well as a single measure of the biomarker not reflecting the longer time period during which the biomarker actually influences the disease. Validity and reliability studies that aim to assess the degree of biomarker error for use of a specific biomarker in epidemiologic studies must be designed to measure *all* of these sources of error; this differs from laboratory validation, which aims to assess only the laboratory component of error. If studies have not been published on these measurement issues, then a validity or reliability study of the biomarker should be conducted to determine its measurement error.

Once the measurement error in the biomarker has been quantified, the researcher can estimate the impact of that magnitude of error on the planned epidemiologic study in terms of the bias in the risk ratio of the relationship of the biomarker to the disease outcome. If there is a large degree of error, the researcher would need to improve the method or select a different one. If the biomarker measure is sufficiently valid to use in an epidemiologic study, then knowing the degree of measurement error will help in interpreting the results.

Definition of terms

The term parent epidemiologic study refers to the epidemiologic study that will use the biomarker. For simplicity, the assumption is made that the parent study is a case–control, cohort, or nested case–control study of the relationship between the biomarker and a binary outcome, such as incident disease or death. Measurement error in the biomarker leads to bias in the risk ratio for the association of the biomarker to disease in the parent study. This bias is called information bias or misclassification bias.

The measurement error for an individual can be defined as the difference between their measured biomarker and true exposure. The true exposure can be conceptualized as the underlying biologic or external factor that the biomarker is meant to measure (the causal factor for etiologic studies), without laboratory or other sources of error. If the biomarker measure can fluctuate over time, the true exposure would also be integrated over the time period of interest (e.g. the average of the true exposure over the etiologically important time period for etiologic studies). Nondifferential measurement error occurs when the measurement error does not differ between the disease and non-disease groups in the parent epidemiologic study. Differential measurement error occurs when the degree of biomarker error differs between those with and without the disease in the parent study. The sources and effects of both differential and nondifferential measurement error will be discussed in this chapter.

Validity is the relation of the biomarker measure to the true exposure in a population of interest. Measures of validity are parameters that describe the measurement error in the population. A validity study is defined here as one in which a sample of individuals is measured twice: once using the biomarker measure of interest and once using a perfect (or near-perfect) measure of the true exposure, and the values compared.

Often a perfect measure of the exposure does not exist or is not feasible to use in a validity study. In a reliability study, repeated measurements of the same biomarker are taken on a group of subjects and compared; they usually only measure part of the measurement error. However, certain designs of reliability studies can be used to measure the validity of a biomarker without having a perfect measure of the biomarker.

Overview of chapter

The first topics covered in this chapter are sources of measurement error in biomarkers and design issues in validity and reliability studies for biomarkers. The chapter then covers the parameters used in a validity study to measure the error in a binary biomarker and in a continuous biomarker. Equations are given for using these parameters to estimate the bias in the risk ratio in an epidemiologic study that will use the biomarker for both binary and continuous biomarkers. While these equations rely on simplifying assumptions, the purpose is to allow the researcher to easily estimate the impact of biomarker error on the parent epidemiologic study. Finally, these same concepts will be addressed for reliability studies.

Techniques to reduce biomarker measurement error, and therefore to reduce the bias in the results of the parent study caused by measurement error, are of great importance. Approaches to reduce measurement error are only briefly mentioned in this chapter, but are covered throughout the book.

Many related topics are beyond the scope of this chapter. The reader is referred to other sources for the effects of measurement error in a categorical measure with more than two categories (1–3), the design and analysis of more complex types of reliability studies (4), and the effect of measurement error when the parent study is the relationship between a biomarker and a continuous outcome (5). General reviews of measurement error effects in epidemiologic studies, and their correction for continuous and/or categorical exposures, are given in (6–12).

Table 8.1. Sources of measurement error in biomarkers in epidemiologic studies

Errors in the choice of laboratory method or specimen (as a measure of the true exposure of interest)
• Method may not measure all sources of the true etiologic exposure of interest (e.g. use of serum beta-carotene when the disease is influenced by all carotenoids)
• Method may measure other related exposures that are not the true exposure of interest (e.g. use of serum total carotenoids, when the disease is only influenced by beta-carotene)
• Biomarker value in tissue sampled may not equal the value in the target tissue
Errors or omissions in the protocol
• Failure to specify the protocol in sufficient detail regarding timing and method of specimen collection, specimen handling and storage procedures
• Failure to specify the laboratory analytic procedures in sufficient detail
• Failure to include standardization of the instrument periodically throughout the data collection
Errors due to variations in execution of the protocol
• Variations in method of specimen collection
• Variations in specimen handling or preparation before reaching the laboratory or freezer
• Variations in length of specimen storage or freeze-thaw cycles (leading to possible analyte degradation)
• Contamination of specimen
• Variations in technique between laboratories
• Variations in technique between laboratory technicians
• Variations between batches (due to different batches of chemicals, drift in calibration of instrument)
• Any of the above that vary between disease and non-disease groups (e.g. unequal assignments of lab technicians to cases and controls)—*differential measurement error*
• Biases due to knowledge of lab technicians of disease status—*differential measurement error*
• Random variation within batch
Errors due to biomarker variability between and within subjects
• Biomarker may be influenced by the disease under study, its pre-clinical effects or its treatment or sequelae - *differential measurement error*[a]
• Short-term variability (hour-to-hour, day-to-day) in biomarker within-subjects due to diurnal variation, posture (sitting versus lying down), time since last meal, time since last exposure to agent of interest in relation to the half-life of the biomarker
• Medium-term variability (month-to-month) within subjects due to, for example, seasonal changes in diet, transient illness
• Long-term change (year-to-year) within subjects due to, for example, purposeful dietary changes over time, changes in occupational exposures
• Lack of variability in biomarker with changes in exposure to agent of interest, due to homeostasis

[a] This is a source of differential error for etiologic studies, not for studies of biomarkers for early detection.

Sources of biomarker error and study designs to measure it

Sources of error in biomarkers

When used in an epidemiologic study, there are numerous sources of measurement error in a biomarker in comparison to the true exposure of interest (5,13). Examples have been discussed in previous chapters and are given here in Table 8.1. Measurement error can be introduced by errors in the choice of laboratory method selected to be a measure of the exposure of interest. The method selected may not measure all sources of the true exposure. For example, if the true exposure of interest is total carotenoids, then using only serum β-carotene as the biomarker would not capture all of the relevant exposure. Alternately, the measure selected may detect other related exposures beyond the etiologically significant one (i.e. it may not be 'specific' to the exposure of interest). For example, if the true exposure of etiologic importance is β-carotene, the choice of serum total carotenoids as the biomarker measure would include other exposures not relevant to the epidemiologic true exposure. Other sources of error that must be considered, especially in the selection of the biomarker method, are whether the method has a sufficiently long half-life in the tissue selected to measure the exposure of interest, and whether homeostasis leads to the method not reflecting the actual level of external exposure. These sources of error can be minimized by careful selection of the biomarker measure to be used in the epidemiologic study.

Further sources of biomarker error are variations in the method of specimen collection and laboratory technique used between subjects. Due to variations in the collection and handling of the specimen by the study field staff, and by variations in the length of time or temperature the specimen was stored before analysis, errors often occur during the conduct of the epidemiologic study. Additional sources of error are the variations that occur between batches and between laboratory technicians even when the protocol is well specified. To reduce these sources of error, the protocol needs specific details in terms of subject instructions (e.g. fasting), method of specimen collection, handling of the specimen (e.g. maximum time at room temperature), methods of specimen processing before storage or analysis, and laboratory procedures. These procedures also must be carefully monitored throughout the study.

A final source of error that is common in molecular epidemiology is due to medium-term (e.g. month-to-month) variability or long-term change in the biomarker over years within-subjects. This type of error is often ignored when assessing laboratory measurement error, but can have great impact on an epidemiologic study. This is due to the fact that unless the biomarker is a fixed characteristic within individuals, the underlying true biomarker (that influences the disease of interest) is rarely an individual's measured biomarker on a single day, but rather the average over some much longer time period. Thus, even a perfect measure of the biomarker at a single point in time could be a poor measure of the true etiologic exposure. For example, even if an ideal laboratory method existed for serum estradiol in women, it could be a very poor measure of the true exposure (e.g. average serum estradiol) over the prior 15 years, which may influence breast cancer. This source of error can be controlled for by averaging multiple measures of exposure collected periodically over the time period of interest (12,14).

While it is essential to minimize the above sources of biomarker error, it is even more important to ensure that any errors are nondifferential between those with and without the disease (or other outcome) under study. Differential measurement error can invalidate a study, as discussed below, and should be avoided. A primary concern in case–control studies of biomarker-disease associations, when the specimens are obtained after diagnosis for the cases, is that the biologic effects of the disease or its treatment may affect the biomarker. In such situations, the biomarker does not measure the true (e.g. long-term, pre-disease) exposure for cases. (This concern is for etiologic studies, not for studies of biomarkers tested for early detection.) Differential measurement error can be avoided or reduced by selecting a cohort study rather than a case–control design when the disease or its treatment can affect the biomarker. The early cases occurring in a cohort study may also have their biomarker influenced by the preclinical phase of the disease under study. However, this can be tested by removing cases with diagnoses that occur within some time period (e.g. a year or two after the specimen collection) to see if this modifies the cohort study results.

Other potential sources of differential biomarker error are the laboratory technicians' knowledge of the disease status of the subjects and differences in the specimen collection, or other methods, between those with and without the disease. Thus, not only must laboratory personnel be blinded to disease status, but also the researcher must ensure that all procedures of specimen collection, processing, storage and analysis are identical for cases and controls. One cannot, for example, collect specimens in a clinic and immediately freeze them for cases, yet collect specimens in the field when freezing will be delayed for controls, if length of time at room temperature can have any impact on the biomarker. Since some variation is inevitable, it is important that the sources of error are matched by case–control status or adjusted for in the analysis. For example, one can control for systematic differences between laboratory technicians, between laboratory batches, or between specimens stored for different lengths of time, through matching controls to cases on these factors (15).

Design of validity and reliability studies

To design a validity or reliability study that measures the amount of measurement error that will occur in the parent epidemiologic study, several design issues must be considered. First, the subjects in the validity or reliability study should represent those in the parent epidemiologic study. The subjects could be a random sample from the parent study (e.g. 100–200 randomly selected individuals from the larger parent study). If that is not possible, the subjects in the validity/reliability study should be comparable to the subjects in the parent study in terms of age, sex and other parameters that could influence the distribution (variance) of the biomarker. Second, the biomarker should be collected, processed, stored and analysed in the validity/reliability study using the same procedures that will be used in the parent study.

A third issue is the selection of the comparison measure to be used in a validity study. Subjects need to be measured using a perfect measure of the true exposure in the validity study to compare to the imperfect biomarker. Sometimes a perfect or near-perfect measure exists that is too costly or not feasible for the parent epidemiologic study, but could be used for a validity study. This true measure must reflect the underlying true exposure without error, including without error due to variation in laboratory procedures or variations over time. The last issue is particularly problematic because the true biomarker of interest is often the average value over many years.

A reliability study can be conducted even when a perfect measure is not feasible or does not exist. For reliability studies, it is ideal if the two or more repeated biomarker measures taken on each subject vary in a way so as to capture all of the sources of error in the biomarker. Therefore, the repeated measures on a subject should be based on specimens taken years apart to reflect error due to year-to-year variation, and be analysed by different laboratory technicians if more than one will be used in the parent study, etc. This differs from a reliability study that aims to assess only the laboratory component of error; those studies might split a single specimen (e.g. blood from a single blood draw) from each subject and send the two samples per subject to the same laboratory for analysis. The importance and methods of measuring all sources of error will be discussed in more detail below.

In addition, the researcher should consider conducting the validity/reliability study on two groups: those with the disease and those without the disease. This is particularly important if there is the possibility that the biomarker error could differ between the disease and non-disease groups that will be used in the parent study. As noted earlier, this is a concern if the parent epidemiologic study is a case-control study in which the disease could influence the biomarker test. For validity/reliability studies to be able to assess differential measurement error in the biomarker measure between cases and controls, the comparison measure must be perfect (i.e. a validity study), or if an imperfect comparison measure is used, it must not have differential error. For example, the comparison measure could be based on specimens collected years before diagnosis for cases and during a similar period for controls. The design, analysis and interpretation of studies to measure differential error are only briefly discussed in this chapter (see (16) for a more detailed review).

Finally, a validity or reliability study should be analysed using parameters that provide information about the impact of biomarker measurement error on the parent epidemiologic study. These parameters, and their interpretations in terms of bias in the risk ratio in the parent study, are discussed in the remainder of this chapter.

Measuring the error in a binary biomarker: Sensitivity and specificity

Binary (dichotomous) biomarkers are those that classify an analyte or characteristic as present (positive) or absent (negative) for each study subject. Measurement error in a binary biomarker is usually referred to as misclassification. Binary biomarkers are subject to all of the sources of measurement error as described above and in Table 8.1.

The degree of misclassification in a binary biomarker is measured by its sensitivity and specificity. These can be measured in a validity study in which the biomarker under evaluation (the mismeasured biomarker) is compared to a perfect measure of the underlying true exposure among a sample of the population of interest. Individuals are then cross-classified by their results on each test:

		True Exposure	
		+	–
Classified by	+	A	B
Biomarker Test	–	C	D

The sensitivity (sens) of the biomarker under evaluation is the proportion of those who are true positives (positive on the criterion test) and are correctly classified as positive by the biomarker test:

$$\text{sens} = \frac{a}{a+c} \ .$$

(Note that the definition given here of sensitivity is different from the meaning in some laboratory contexts, i.e. the lowest level detectable by a measurement method.) The specificity (spec) is the proportion of those who are true negatives and are classified as negative by the biomarker test:

$$\text{spec} = \frac{d}{b+d} \ .$$

Even though both sensitivity and specificity can range from zero to one, it is assumed that sensitivity plus specificity is greater than or equal to one. In other words, for the biomarker test to be considered a measure of the true biomarker, the probability that the biomarker test classifies a truly positive person as positive (sensitivity) should be greater than the probability that it classifies a truly negative individual as positive (1 – specificity) (i.e. sens > 1 – spec, or sens + spec > 1). Thus, the parameter (sens + spec –1), called the Youden misclassification index (17), is a good measure of the total degree of misclassification. If the Youden index is close to 1, the biomarker test is close to perfect, and if the Youden index is close to zero, the biomarker has little association with the true exposure.

For a validity study to measure sensitivity and specificity of a biomarker, the study sample may be subjects selected independent of their biomarker status, or who are true positives and those who are true negatives by the perfect test. However, one cannot sample subjects based on the results of the mismeasured biomarker test and correctly compute sensitivity and specificity.

If the validity study is conducted on a sample of cases and a sample of controls, then sensitivity and specificity would be computed separately for the cases and for the controls.

Impact of error in a binary biomarker on epidemiologic studies

Effect of nondifferential misclassification on the odds ratio

The effects of misclassification of a binary biomarker on the results of the parent epidemiologic study of the relationship between the marker and a binary disease are straightforward (18–24). In an unmatched case–control study of a binary biomarker, under the assumption that the disease is correctly classified, the effect of misclassification of the biomarker is to rearrange individuals in the true 2x2 table into an observable 2x2 table. Individuals in the disease group remain in the disease group, but may be misclassified as to biomarker status, and the non-disease group is also rearranged as to biomarker status:

True Classification
Disease

		+	–
True Exposure	+	P_1	P_2
	–	$1 – P_1$	$1 – P_2$

Odds ratio : $OR_T = \dfrac{P_1(1 - P_2)}{P_2(1 - P_1)}$

ObservableClassification
(Misclassification)
Disease

		+	–
Biomarker Test	+	p_1	p_2
	–	$1 – p_1$	$1 – p_2$

$OR_O = \dfrac{p_1(1 - p_2)}{p_2(1 - p_1)}$

P_1 and P_2 are the true proportions of subjects who are exposure-positive in the disease and non-disease groups respectively, and similarly p_1 and p_2 refer to the proportions that would be “observable” as positive by the biomarker test in the two groups. The term observable means what would be seen, on average, when the imperfect biomarker is used in the parent epidemiologic study.

There is nondifferential misclassification when the sensitivity of the biomarker test is the same for both disease and non-disease groups in the parent study, and the specificity is the same for both groups. The misclassification leads the observable p_1 and p_2 to be different from the true P_1 and P_2 (21):

$$\left.\begin{array}{l} p_1 = \text{sens} \bullet P_1 + (1 - \text{spec}) \bullet (1\text{-}P_1) \\ p_2 = \text{sens} \bullet P_2 + (1 - \text{spec}) \bullet (1\text{-}P_2). \end{array}\right\}(1)$$

The first equation states that the proportion of cases who will be classified by the biomarker test as positive (p_1) is made up of a proportion (sens) of those truly exposed (P_1) in the disease group, plus a proportion (1-spec) of those truly unexposed (1-P_1) in the disease group. The second equation expresses the same concept for the non-disease group.

The association between the biomarker and disease in the parent study would be measured by the odds ratio in a case–control or nested case–control study. (The results presented here would be similar for the hazard ratio from a cohort study as well.) When there is measurement error, the observable odds ratio, OR_O, differs from the true odds ratio, OR_T, because OR_O is based on p_1 and p_2:

$$OR_O = \frac{p_1(1 - p_2)}{p_2(1 - p_1)}\ . \qquad (2)$$

The magnitude of the bias can be estimated by computing p_1 and p_2 from Equation 1 and OR_O from Equation 2 and comparing it to OR_T, using estimates of P_1 and P_2.

As an example, suppose current infection with *Helicobacter pylori* (*H. pylori*), an organism associated with several upper digestive tract diseases, is the true exposure of interest in a cohort study being planned. An ELISA test on serum, although imperfect, is the most feasible exposure measure to be used in the epidemiologic study. Information on the accuracy of the ELISA test comes from a validity study conducted in Taiwan, China on 170 patients undergoing gastroendoscopy (25). The serum ELISA test was compared to a highly accurate measure, assessed by either a positive culture or two positive tests among three others (histology, Campylobacter-Like Organism (CLO) test and 13C-urea breath test), with these results:

		True Exposure	
		H.pylori +	*H.pylori* –
ELISA Test	*H.pylori* +	103	16
	H.pylori –	4	47
	Total	107	63

Sensitivity and specificity of the ELISA test were:

sens = 103/107 = 0.96
spec = 47/63 = 0.75.

These estimates can be used to approximate the effect of the biomarker error on the results of the epidemiologic study. If one assumes that the measurement error in the future cohort study will be nondifferential, i.e. that the sensitivity and specificity are the same for cases and controls, then Equations 1 and 2 can be used to estimate the observed odds ratio, OR_O. If the estimated true *H. pylori* infection prevalence is 70% in cases (P_1) and 35% in controls (P_2), then the true odds ratio, OR_T, is:

$$OR_T = \frac{.70(.65)}{.35(.30)} = 4.3$$

Then by Equations 1 and 2:

$$p_1 = 0.96 \bullet 0.70 + 0.25 \bullet 0.30 = 0.75$$
$$p_2 = 0.96 \bullet 0.35 + 0.25 \bullet 0.65 = 0.50$$
$$OR_O = \frac{.75(.50)}{.50(.25)} = 3.0$$

This validity study shows that a study using the misclassified *H. pylori* test would find 75% of cases positive and 50% of controls positive (rather than 70% and 35%, respectively, as the true probabilities of exposure), and would yield an observed odds ratio of 3.0 rather than the true odds ratio of 4.3.

Nondifferential misclassification leads to an attenuation of the odds ratio towards the null hypothesis value of 1 (20). The observable odds ratio does not cross over the null value of 1, under the reasonable situation that sens + spec > 1 (see above). The degree of attenuation in the observable odds ratio depends on the sensitivity and specificity of the biomarker test and on P_1 and P_2, or equivalently, on the true odds ratio and on P_2, the proportion of the non-disease group who are truly exposed. Table 8.2 gives further examples of the effect of nondifferential misclassification on the odds ratio for reasonable values of sensitivity (0.5–0.9), specificity (0.8–0.99), P_2 (0.1, 0.5), and a true odds ratio of 2 and 4. As can be seen from the table, the attenuation in the odds ratio can be considerable. When the proportion who are truly exposed in the non-disease group is low (e.g. P_2 = 0.1 in upper half of Table 8.2), the attenuation of the odds ratio is severe except when the specificity is very high (e.g. spec = 0.99). When the proportion who are truly exposed is high (e.g. P_2 = 0.5 in lower half of Table 8.2), the observed OR is strongly attenuated from the true OR except when the sensitivity is very high (e.g. sens = 0.9). Even strong associations between the true biomarker and disease would be obscured by moderate values of sensitivity and specificity. For example, for sens = 0.7, spec = 0.8,

Table 8.2. Impact of nondifferential misclassification of a binary biomarker on the Observable Odds Ratio (OR_O)

Biomarker Test Sensitivity	Biomarker Test Specificity	True OR=2.0 OR_o[b]	True OR=4.0 OR_o[b]
P_2 = 0.1[a]			
0.5	0.80	1.14	1.38
0.7	0.80	1.23	1.64
0.9	0.80	1.32	1.92
0.5	0.90	1.28	1.76
0.7	0.90	1.39	2.09
0.9	0.90	1.48	2.41
0.5	0.99	1.75	3.06
0.7	0.99	1.83	3.33
0.9	0.99	1.89	3.61
P_2 = 0.5[a]			
0.5	0.80	1.24	1.46
0.7	0.80	1.40	1.83
0.9	0.80	1.64	2.59
0.5	0.90	1.35	1.69
0.7	0.90	1.50	2.07
0.9	0.90	1.73	2.85
0.5	0.99	1.48	1.96
0.7	0.99	1.61	2.33
0.9	0.99	1.82	3.11

[a] P_2 is the proportion with the true exposure in the non-disease group. By definition of OR_T, P_1, the proportion with the true exposure in the diseased group is: $P_1 = P_2 \bullet OR_T/(1 + P_2\,(OR_T - 1))$.
[b] OR_O from Equations 1 and 2.

and $OR_T = 4.0$, the observable odds ratio would be 1.64 for $P_2 = 0.01$ and 1.83 for $P_2 = 0.5$. These observable odds ratios would not be detectable as different from the null value of 1 unless the parent epidemiologic study sample size was large.

Effect of differential misclassification on the odds ratio

There is differential misclassification when the sensitivity of the biomarker test for the disease group differs from that for the non-disease group, and/or the specificity of the biomarker test for the disease group differs from that for the non-disease group. Differential misclassification can have any effect on the odds ratio: the observable odds ratio can be closer to the null hypothesis of OR = 1, be further from the null, or crossover the null compared with the true odds ratio. Thus, while the odds ratio under nondifferential measurement error can be assumed to be conservative (biased towards the null), the odds ratio when there is nondifferential error could be biased away from the null or even be in the wrong direction (e.g. it could make the biomarker appear to be a risk factor when it is, in fact, a protective factor). Equations 1 and 2 can be used to estimate the impact of differential measurement error, by using the estimates of sensitivity and specificity in the disease group for the equation for p_1, and estimates of sensitivity and specificity in the non-disease group for the equation for p_2.

Measuring the error in a continuous biomarker using a validity study

Often a biomarker assay yields quantitative information about the amount of an analyte in a biologic specimen; these measures can usually be considered to be continuous variables. This section covers the parameters used to assess measurement error in a continuous biomarker from a validity study in which each subject is measured twice: once using the mismeasured biomarker and once using a perfect measure of the true exposure of interest.

The theory of measurement error in continuous variables was developed in the fields of psychometrics, survey research and statistics (26–32). The effects of measurement error also have been derived in the context of epidemiologic studies of a continuous exposure variable and a binary disease outcome (3,33–35).

A model of measurement error

A simple model of measurement error in a continuous measure X is:

$$X_i = T_i + b + E_i,$$

where:

$\mu_E = 0$
$\rho_{TE} = 0.$

In this model for a given individual i, the measured biomarker, X_i, differs from the true value, T_i, by two types of measurement error. The first is systematic error, or bias, b, that would occur (on average) for all measured subjects. The second, E_i, is the additional error in X_i for subject i. E will be referred to as subject error to indicate that it varies from subject to subject. It does not refer just to error due to subject characteristics; rather it includes all of the sources of error outlined in Table 8.1.

For the population of potential study subjects, X, T and E are variables with distributions (e.g. the distribution of E is the distribution of subject measurement errors in the population of interest). X, T and E would have expected means in the population of interest denoted by μ_X, μ_T, and μ_E, respectively, and variances denoted by σ^2_X, σ^2_T, and σ^2_E. Because the average measurement error in X in the population is expressed as a constant, b, it follows that μ_E, the population mean of the subject error, is zero. The assumption of the model that the correlation coefficient of T with E, ρ_{TE}, is zero states that the true value of the biomarker is not correlated with the measurement error. In other words, individuals with high true values are assumed to not have systematically higher (or lower) errors than individuals with lower true values.

Measures of measurement error: Bias and validity coefficient

Two measures of measurement error are used to describe the relationship of X to T in the population of interest, based on the above model and assumptions. One is the bias (i.e. the average measurement error in the population):

$$b = \mu_X - \mu_T .$$

The bias in X can be estimated from a validity study as the difference between the mean of X and the mean of T: $b = \overline{X} - \overline{T}$. If b is positive, then X overestimates T on average; if b is negative, then X underestimates T on average.

The other measurement error is a measure of the precision of X (i.e. the variation of the measurement error in the population). One measure of precision is σ^2_E, the variance of E, which is often called the within-subject variance. (Note that b is a constant for all subjects and therefore does not contribute to the variance of the error, σ^2_E.) A more useful measure of precision is the correlation of T with X, ρ_{TX}, termed here as the validity coefficient of X. The measure ρ_{TX} is important because it relates the within-subject variance, σ^2_E, to the total variance, σ^2_X, and it is this ratio, along with the bias, that measures the impact of biomarker error on the parent epidemiologic study. ρ_{TX} can range between zero and one, with a value of one indicating that X is a perfectly precise measure of T. ρ_{TX} is assumed to be zero or greater (i.e. for X to be considered to be a measure of T, X must be at a minimum positively correlated to T).

The validity coefficient ρ_{TX} can be estimated in a validity study by the Pearson correlation coefficient of X with T. Thus, using the standard interpretation of a correlation coefficient, the correlation squared (ρ^2_{TX}) can be interpreted as the proportion of the variance of X explained by T. For example, if ρ_{TX} were 0.8, this would mean that only 64% of the variance in X is explained by T, with the remainder of the variance due to the error.

To further understand the concepts of bias and precision, consider a situation in which X has a systematic bias, but is perfectly precise (i.e. $E_i = 0$ for all subjects). Suppose that the only source of error in a measure of serum cholesterol, for example, were that it quantified each individual exactly 100 mg/dl too high. X would be biased (b = 100 mg/dl), but would have perfect precision ($\rho_{TX} = 1.0$). Then, in a population, the measure X, even though it has systematic measurement error, could be used to perfectly order each person in the population by their value of T.

There could also be situations in which there is no bias, yet there is lack of precision. Suppose that the measurement error, E_i, varied from person-to-person, but for some subjects their measured X was higher than their actual T, and for other subjects their measured X was lower than their actual T, but X on average in the population equaled the average T in the population. In this situation there is no bias (b = 0), but there is lack of precision (ρ_{TX} < 1.0). In this case, the ordering of subjects is lost. Of course, most likely a biomarker has both bias and lack of precision.

The degree of measurement error is not an inherent property of a biomarker test, but rather is a property of the test applied using a particular protocol to a specific population. Therefore, the error can vary for a biomarker test when applied using a different protocol or when applied to different population groups. In addition, ρ_{TX} is generally greater in populations with greater variance of the true exposure (36). Therefore, a validity study conducted on one population may not directly apply to another study population.

Finally, measurement error could differ between those with and without the disease, particularly in a case–control study. Separate assessment of the bias and precision in these two groups is needed to assess differential error (see below).

The terminology surrounding measurement error varies between fields. In this chapter, the terms validity, accuracy and measurement error are used as general terms reflecting the relationship of X to T, including both the concepts of bias and precision. (In laboratory quality control, the terms validity and accuracy are sometimes used to refer to unbiasedness only.)

Impact of error in a continuous biomarker on epidemiologic studies

When the bias and validity coefficient of the biomarker (X) are known, one can estimate the impact of the degree of measurement error in X on the parent epidemiologic study that will use X. Both nondifferential and differential measurement errors will be discussed, but first the effect of measurement error on a single study population will be considered.

Effect of measurement error on the observable mean and variance

In a single study population, both the mean and variance of the measured biomarker X would differ from the true mean and variance due to measurement error. Under the above model, the population mean of X would differ from the true mean (the population mean of T) by b:

$$\mu_X = \mu_T + b. \qquad (3)$$

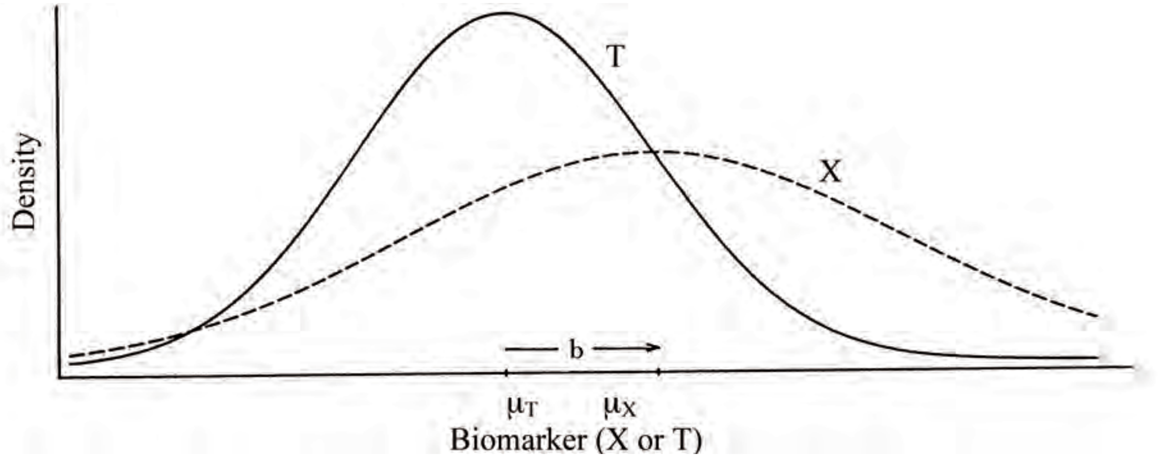
The population variance of X, based on the above model, would be (30):

$$\sigma^2_X = \sigma^2_T/\rho^2_{TX}. \qquad (4)$$

The variance of X in the population is greater than the variance of T, due to the addition of the variance of the measurement error. For example, if the validity coefficient (ρ_{TX}) were 0.8, then the variance of X would be 56% greater than the variance of T ($\sigma^2_X = \sigma^2_T/.8^2 = 1.56\ \sigma^2_T$ by Equation 4).

Figure 8.1 demonstrates the effect of measurement error on the distribution of X in a population using a normally distributed biomarker and normally distributed error as an example. The bias in the measure X causes a shift in the distribution of

Figure 8.1. Distribution of true (T) and measured (X) biomarker



X compared with T. The increased variance of X compared with T (measured by ρ_{TX}) causes a flattening of the distribution of X. Even if a measure X were correct on average (b = 0), there could still be substantial measurement error due to lack of precision, which could lead to a greater dispersion in the measured exposures.

Effect of nondifferential measurement error on the odds ratio

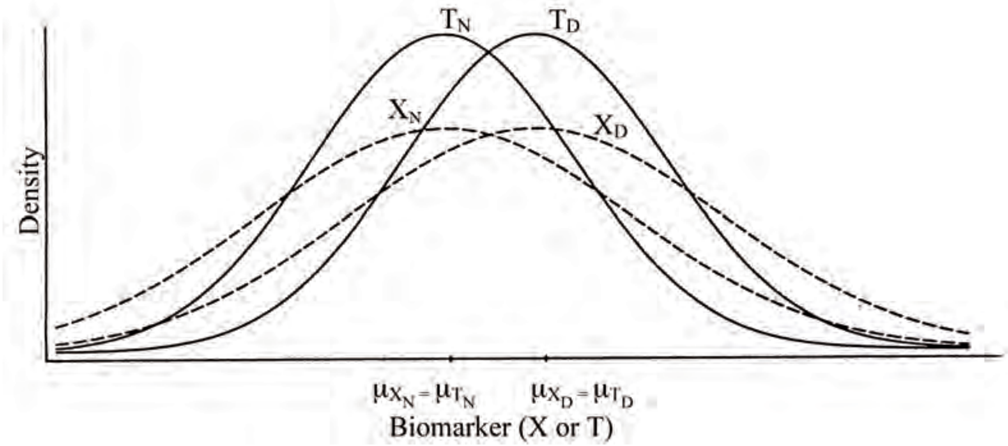
While measurement error has an effect on the observable mean and variance of an exposure variable within a single population, a greater concern would be the impact of measurement error in an epidemiologic study comparing those who have the disease of interest to those who do not. In a case–control or nested case–control study, the common measure of association between a biomarker and disease is the odds ratio, which is often expressed as the odds ratio of disease for a *u* unit increase in the level of the biomarker. The results given here also approximately apply to estimates of the hazard ratio from data from a cohort study and the risk ratio from a matched case–control study (33). The results given in this section do not apply to odds ratios expressed as odds of disease for the upper quantile of the biomarker versus lowest quantile. They also do not apply when X and T are measured in different units in the validity study. For a discussion of these situations, see (12).

Errors in the measurement of the biomarker X would bias the odds ratio in the epidemiologic study. There is nondifferential misclassification when there is equal bias and equal precision (equal ρ_{TX}) in the biomarker test when applied to both the disease and non-disease groups in the parent epidemiologic study. Figure 8.2 illustrates the effects of nondifferential misclassification. Under nondifferential misclassification, the distribution of exposure in cases and in controls may shift, but because there is equal bias for the two groups, they are not shifted with respect to each other. However, the lack of precision flattens and leads to more overlap and less distinction between the distributions of X_D, the biomarker in the disease group, and of X_N, the biomarker in the non-disease group, compared with the distributions of the true exposure in the two groups (T_D and T_N).

The effect of nondifferential measurement error in X on the odds ratio can only be easily quantified when certain simplifying assumptions are made. Results can be derived for case–control studies under the following assumptions: a) X_D and X_N meet the assumptions of the simple model of measurement error given above; b) T_D and T_N are normally distributed with different true means in the disease and non-disease groups, but the same variance, σ^2_T ; c) the bias in X is the same for the two groups; and d) the errors, E, are normally distributed with mean zero and common variance,σ^2_E, in the two groups. Assumption c and the second part of assumption d are the assumptions of nondifferential error (i.e. equal bias and equal precision of X_D and X_N).

The above assumptions imply a logistic regression model for the probability of disease (pr(d)) as a function of true biomarker T, with a true logistic regression coefficient β_T (37):

Figure 8.2. Effect of nondifferential measurement error (equal bias and precision) on distribution of true (T) versus measured (X) biomarker in an epidemiologic study comparing disease (D) and non-disease (N) groups



$$\log \frac{pr(d)}{1\text{-}pr(d)} = a_T + b_T T.$$

The true odds ratio for any *u* unit increase in T would be $OR_T = e^{b_T u}$.

With measurement error in the biomarker test X, the assumptions also lead to a logistic model:

$$\log \frac{pr(d)}{1\text{-}pr(d)} = a_o + b_o X.$$

$OR_O = e^{b_O u}$ is the observable odds ratio for a *u* unit increase in X.

The observable logistic regression coefficient, β_O, differs from β_T due to the measurement error in X. Under nondifferential misclassification, β_O is attenuated towards the null value of zero in comparison to β_T (34,37) by this equation:

$$\beta_O = \rho^2_{TX}\beta_T\,. \qquad (5)$$

Equivalently, OR_O is attenuated towards the null value of 1 in comparison to OR_T by this equation:

$$OR_O = OR_T^{\rho^2_{TX}}\,. \qquad (6)$$

This states that the observable odds ratio for any fixed difference in units of the biomarker is equal to the true odds ratio for the same fixed difference to the power ρ^2_{TX}. Since $0 \leq \rho^2_{TX} \leq 1$, the observable odds ratio will be closer to the null value of 1 (no association) than the true odds ratio. The observable odds ratio does not cross over the null value if X and T are at a minimum positively correlated.

Equation 6 shows that the attenuation in the odds ratio under nondifferential misclassification is a function of the precision of X (measured by ρ_{TX}), but *not* of the bias in X. Thus, even when X is correct on average for cases and correct on average for controls (bias = 0 for cases and for controls), the lack of precision of X can substantially bias the odds ratio. Examples of the effects of nondifferential measurement error in a biomarker on the odds ratio, based on the assumptions above and Equation 6, are given in Table 8.3. The table shows that biomarkers with a validity coefficient ρ_{TX} of 0.5 would obscure all but the strongest associations. For example, when $\rho_{TX} = 0.5$ and the true odds ratio for a *u* unit change in the biomarker was 4.0, this odds ratio would be attenuated to an observed odds ratio of 1.41. Furthermore, measures as precise as $\rho_{TX} = 0.9$ still lead to a modest attenuation (e.g. a true odds ratio of 4.0 would be attenuated to 3.07).

Effect of nondifferential measurement error on power and sample size

The examples above show that nondifferential measurement error in a biomarker leads to attenuation of the odds ratio for the association of the biomarker with the disease. This attenuation of the odds ratio would reduce the power of the epidemiologic study that uses the biomarker if the sample size were fixed. If the sample size was not fixed, it would lead to a need for a larger sample size to detect the attenuated odds ratio as different from the null value of 1.

When a continuous exposure with measurement error is used in an epidemiologic study, the sample size needed, N_X, is compared to the sample size needed in a study in which the exposure is measured without error, N_T. A simple formula shows this comparison (14,38):

$$N_X = N_T / \rho^2_{TX}.$$

This formula may be of theoretical interest only, since estimates of the parameters needed when calculating the required study sample size should be based on the mismeasured exposure (e.g. σ^2_X); as these estimates are usually available, the sample size calculations will yield the correct N. However, the above equation can be used to show the potentially dramatic effects of inaccurate biomarker measurement on the sample size required. For example, if the correlation between T and X is 0.7 ($\rho^2_{TX} = 0.49$), then the sample size required when the imperfect measure is used is twice that required if a perfect measure were available. This shows that the error in biomarkers, with what is considered to be a good validity coefficient, still leads to a large increase in required sample size for the epidemiologic study that will use the biomarker.

Effect of differential measurement error on the odds ratio

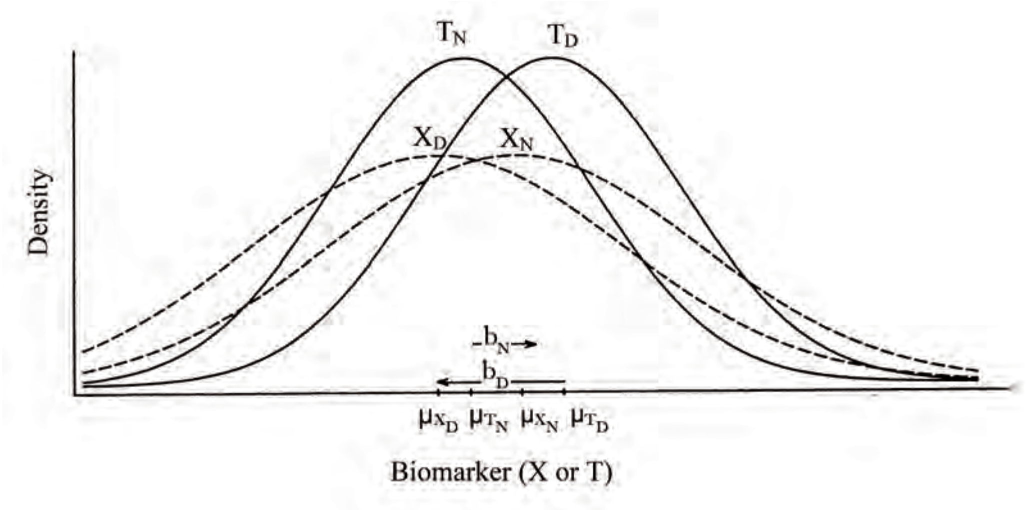
Differential measurement error occurs when the bias in the mismeasured biomarker in the disease group differs from the bias in the non-disease group, and/or the precision differs between groups. As noted above, differential measurement error should be a concern in a case–control study when the biomarker is measured within the preclinical disease phase before diagnosis or anytime after diagnosis, unless the marker is fixed (e.g. genotype). Differential bias has the most problematic effects: depending on the magnitude and the direction of the biases in X_D and X_N, the observable odds ratio for any *u* unit increase in X, $OR_O = e^{b_O u}$, could be towards the null value of one, away from the null, or cross over the null value compared with the true odds ratio.

Figure 8.3 presents a graphical example of differential measurement error, in particular, differential bias between cases and controls. In the figure, the true mean exposure level in the disease group, μ_{TD}, is greater than the true mean level in the non-disease group, μ_{TN}. This would lead to an odds ratio above 1 for any *u* unit increase in T. In this example, the bias for the non-disease group is positive, so the distribution of X_N is shifted to the right relative to T_N, and the bias among those with disease is negative, so the distribution of X_D is shifted to the left relative to T_D. Differential bias would cause the observable odds ratio to cross over the null value of one (i.e. the biomarker would appear to be a protective factor, rather than a risk factor, as the controls would appear to have higher mean exposure than the cases).

Differential bias could be assessed by comparing the bias ($\overline{X} - \overline{T}$) for cases with the bias ($\overline{X} - \overline{T}$) among controls. To assess differential bias, T does not need to be perfect; rather, T only needs to have nondifferential bias (e.g. T could be based on specimens drawn years before diagnosis).

An example of differential bias comes from several case–control studies which found that low serum cholesterol, measured at the time of diagnosis, was a risk factor for colon cancer (39). This could be an artefact if increased catabolism, or other effects of colon cancer, reduce serum cholesterol. In fact, it was found that serum cholesterol

Figure 8.3. Effect of differential measurement error ($b_1 \neq b_2$) on distribution of true (T) versus measured (X) biomarker in an epidemiologic study comparing disease (D) and non-disease (N) groups



measured 10 years before diagnosis was higher in colon cancer cases than controls (40). This suggests that serum cholesterol measured at the time of diagnosis had differential bias; it likely underestimated the true etiologic exposure (say, true serum cholesterols a decade before) among cases due to the effects of the cancer, while it may have overestimated the true serum cholesterol a decade before among the controls (due to a tendency of cholesterol levels to increase with age).

Differential bias is a greater concern than differential precision because, as described above, differential bias can lead to a shift in the distribution of the biomarker in one group relative to the other. Differential measurement error will also occur if precision differs between groups. If there were no differential bias, but precision differed, the shape of the odds ratio function could change. For example, the observable odds ratio curve could be U-shaped when the true exposure–disease relationship is increasing (41).

More details about the design, analysis and interpretation of validity or reliability studies to assess differential measurement error are given in (16).

Measuring the error in a biomarker using a reliability study

The term reliability is used to refer to the reproducibility of a measure, that is, how consistently a measurement can be repeated on the same subjects. Reliability can be assessed in several ways, but only one type, intramethod reliability, will be covered in this chapter. Intramethod reliability studies measure the reproducibility of a measurement on the same subjects repeated two or more times using the same method, but often with some variation. For example, a comparison could be made of a biomarker from a single specimen from each subject analysed by the same laboratory technician twice, or analysed by two laboratory technicians, or from two specimens on each subject collected at two points in time. Reliability studies, in which two different analytic methods are compared, with one better than the other but neither perfect (intermethod reliability studies), are not covered here (see (12)). Measures of reliability are primarily important for what they reveal about the validity of a biomarker test, because as shown above, the bias in the odds ratio in the parent epidemiologic study is a function of the validity of the biomarker measure.

This section covers the parameters used to measure reliability, the interpretation of measures of reliability in terms of measures of validity, and the use of parameters from reliability studies to estimate the bias in the odds ratio in the parent epidemiologic study that will use the biomarker.

A model of reliability and measures of reliability for continuous biomarkers

Suppose each person in a sample of interest is measured two or more times using the same continuous biomarker test that will be used in the parent study. For a given subject i, two (or more) biomarker measurements, X_{i1} and X_{i2}, are obtained. The simple measurement error model described above applies to each measure:

$$X_{i1} = T_i + b_1 + E_{i1}$$

$$X_{i2} = T_i + b_2 + E_{i2}$$

Both X_{i1} and X_{i2} are measures of the subject's true biomarker T_i, but with different errors. In a reliability study, information is available on X_1 and X_2 for each subject, but not on T. A reliability study can yield estimates of the mean of X_1 and X_2 (μ_{X_1} and μ_{X_2}) and of the correlation between the two measures, ρ_X, termed the reliability coefficient.

The intraclass correlation coefficient (ICC) is generally used as the reliability coefficient for continuous biomarkers (see (12,14) for computational formulas). The ICC differs from the Pearson correlation coefficient in that it includes any systematic difference between X_1 and X_2 (i.e. any difference between b_1 and b_2) as part of the subject error E (the error that varies from subject-to-subject). The assumption is that in the parent epidemiologic study, each subject will be measured once, by either X_1 or X_2 (e.g. either by laboratory technician 1 or 2). Therefore, any systematic difference between X_1 and X_2 would not be a systematic bias affecting everyone in the parent study, but would vary between subjects because some are measured by X_1 and some by X_2. Thus, the ICC is equal to 1 only when there is exact agreement on all measures on each subject (which differs from the Pearson correlation coefficient, which is equal to 1 when X_1 is a linear function of X_2). Because X_1 and X_2 will be used as interchangeable measures of X in the parent study, and more than two replicates per subject can be used to compute the ICC, the reliability coefficient of X is written as ρ_X.

Two measures of the validity of a continuous biomarker measure X, the bias and the validity coefficient, were shown to be important in assessing the impact of measurement error on the parent epidemiologic study, which will use X. Unfortunately, reliability studies generally cannot provide information on the bias in X, because a similar bias often affects both X_1 and X_2. The inability of many reliability study designs to yield information on the bias, and on differential bias between cases and controls, is a major limitation. It should be recalled, however, that under nondifferential measurement error (and certain other assumptions), the attenuation of the odds ratio depends only on the validity coefficient and not on the bias. The reliability coefficient does provide information about the validity coefficient, and thus can be used to estimate the effects of measurement error on the parent study under the assumption of nondifferential measurement error.

Relation of reliability to validity under the parallel test model

When certain assumptions are met, reliability studies can yield estimates of the validity coefficient. One such set of assumptions is the model of parallel tests (27,29–31). The first assumption of the parallel test model is that the error variables, E_1 and E_2, are not correlated with the true value T. The second is that E_1 and E_2 have equal variance σ^2_E. This also implies that X_1 and X_2 have equal variance and that X_1 and X_2 are equally precise ($\rho_{TX_1} = \rho_{TX_2}$). This is usually a reasonable assumption in intramethod reliability studies, since X_1 and X_2 are measurements from the same method. Third, it is assumed that E_1 is not correlated with E_2. This important (and restrictive) assumption implies, for example, that an individual who has a positive error, E_1, on the first measurement is equally likely to have a positive or a negative error, E_2, on the second measurement. These assumptions are often summarized by saying that two measures are parallel measures of T if their errors are equal and uncorrelated.

Under the assumptions of parallel tests, it can be shown that (30):

$$\rho_{TX} = \sqrt{\rho_X} \,. \tag{7}$$

This equation states that the validity coefficient of X, ρ_{TX}, can be estimated to be the square root of the reliability coefficient, ρ_X. This result is important because it shows that if the assumptions are correct, the reliability coefficient, which is a measure of the correlation between two imperfect measures, can be used to estimate the correlation between T and X without having a perfect measure of T. The correlation of the replicates of X is less than the correlation of X with T, as each replicate has measurement error.

A reliability study of a biomarker test can often be assumed to have equal and uncorrelated errors if the replicates within each person are sampled over the entire time period to which the true biomarker is intended to relate (if the biomarker can vary over time); the specimen handling, storage and analytic techniques vary in the reliability study as they will in the parent study; and the true exposure is defined as the mean measure over the relevant time period of repeated measures of the assay.

An example comes from a study which examined the reliability of serum hormone levels in premenopausal women (42). The goal was to understand whether a single blood draw was sufficiently accurate to be used in a large prospective study of serum hormones and cancer risk among premenopausal women. The reliability study included 113 women who had blood drawn once a year for three years during both the middle of the follicular and luteal phases of their menstrual cycles. The reliability coefficient (intraclass correlation coefficient) was 0.38 for total estradiol during the follicular phase and 0.45 during the luteal phase. The repeated measures in this study are close to the parallel test model: the errors on each of the repeated measures can be assumed to be equal because the same test procedure was repeated, and the errors are likely to be uncorrelated (i.e. a woman whose hormone measure was higher than her "true" three-year average on one measure is not more likely to be higher than her true average on another measure). This study also measured most sources of error, such as error due to variations in blood processing, storage and laboratory technique, and error due to long-term variation of plasma hormones within women. Therefore, the estimated validity coefficient (ρ_{TX}) for a single measure of total estradiol (X) as a measure of average estradiol over three years (T), based on Equation 7, is 0.62 if blood were drawn during the mid-follicular phase and 0.67 if drawn during the mid-luteal phase.

Based on Equation 7, the results in the last section on the effects of measurement error on the odds ratio can be expressed in terms of ρ_X rather than ρ^2_{TX}. When the parallel model holds, Equation 6 can be written:

$$OR_O = OR_T^{\rho_X} \cdot \tag{8}$$

From the example above (42), use of a single measure of total estradiol during the mid-luteal phase, with a reliability coefficient of 0.45 in a cohort study of total estradiol and breast cancer, would attenuate a true odds ratio of, for example, 4.0 to an observed odds ratio of 1.9 (from Equation 8). Other examples of the bias in the odds ratio (under the parallel test model)

Table 8.3. Impact of nondifferential measurement error in a normally distributed biomarker X on the Observable Odds Ratio (OR_O)

ρ_{TX}[a]	ρ_X[b]	True OR[c]=2.0 OR_O[d]	True OR[c]=4.0 OR_O[d]
.50	.25	1.19	1.41
.60	.36	1.28	1.65
.70	.49	1.40	1.97
.75	.56	1.48	2.18
.80	.64	1.56	2.42
.85	.72	1.65	2.72
.90	.81	1.75	3.07
.95	.90	1.87	3.49

[a] ρ_{TX} is the validity coefficient of X.
[b] ρ_X is the reliability coefficient of X under the parallel test model (see text):

$$\rho_X = \rho^2_{TX}$$

[c] The true OR = is the odds ratio for a *u* unit difference in T for comparison to OR_O for a *u* unit difference in X.
[d] OR_O from Equation 6 or 8. See text for model and assumptions.

from various degrees of unreliability are given in Table 8.3.

Relation of reliability to validity for correlated errors

In actual reliability studies, the assumptions of parallel tests are often incorrect. One assumption that is frequently violated is that of uncorrelated errors. Often the error in one measure is positively correlated with the error in the other ($\rho_{E_1E_2} > 0$). Correlated errors occur when the sources of error in the first measurement on a subject tend to repeat themselves in the second. If in a reliability study, for instance, blood was drawn only once on each subject and analysed twice, and the true marker of interest was the mean value of the biomarker over several years surrounding the time of measurement, there would be correlated error. For example, suppose the reliability of serum β-carotene was assessed by repeated laboratory analysis from a single blood draw. This would have correlated error, as an individual whose β-carotene level was higher on the first measure than the true long-term average (perhaps due to a seasonal variation in intake of β-carotene) would also likely be higher on the second measure than the true value, because the second measure used the same specimen. Because part of the error is repeated in both X_1 and X_2, the errors are correlated. This means the reliability study does not capture all sources of error in X, and therefore the reliability coefficient, ρ_X, is artificially too high.

When the errors of the measures used in a reliability study are positively correlated, then the reliability study can only yield an upper limit for the validity coefficient. Specifically, when X_1 and X_2 are equally precise, and the assumptions of the above model hold except that the errors are correlated, then the validity coefficient is less than the square root of the reliability coefficient (1):

$$\rho_{TX} < \sqrt{\rho_X}\ . \qquad (9)$$

Thus, a measure can appear to be reliable (repeatable) even if it has poor validity. While a low reliability coefficient implies poor validity, a high reliability does not necessarily imply a high validity coefficient. The high reliability may be due instead to correlated errors within subjects. The reliability coefficient is only diminished by part of the error in X (the part that is not repeated in X_1 and X_2), whereas the validity coefficient is a measure of all sources of error. When there is correlated error (i.e. when only part of the error is measured by a reliability study), then the attenuation of the odds ratio will be even greater than that predicted by Equation 8. Reliability studies should be designed, therefore, to capture all of the sources of error in the biomarker X, including error due to variations of specimen collection, variations between laboratory technicians, and within-person variations over time. (The concepts of correlated error, repeated within-person error, and failure of a reliability study to capture all sources of error, each describe the same phenomenon.)

Sometimes it is only possible or desirable to assess some components of error. To assess the laboratory error, a blinded test-retest reliability study on split samples from a single specimen from each subject in the reliability study, analysed in separate batches and by different laboratory technicians (if multiple laboratory technicians were going to be used in the parent epidemiologic study), would yield an intraclass correlation coefficient that measures the laboratory component of error only. Similarly, other reliability studies could be designed to test the effect of handling, storage, and short-, medium- and long-term biologic variation by only varying these components. When only some components of error are measured, there is correlation error and the resulting intraclass correlation just provides an upper estimate of the validity coefficient (see Equation 9). However, by estimating the components of error, the researcher can seek to improve those aspects having the most adverse effects. For example, enhanced laboratory quality control procedures could be used to reduce laboratory error, or multiple specimens (over time) per subject could be used to reduce the error due to medium- or long-term biologic variation. Finally, nested reliability study designs can be used to estimate the different components of error within one reliability study (4).

Coefficient of variation

One additional analytic technique for reliability studies of continuous biomarkers, the coefficient of variation (CV) deserves mention (43). For laboratory measures, reliability is often assessed by repeated analysis of a single reference material. For example, a single pooled blood sample might be analysed 10 times to yield measures of the biomarker X. In such studies, the mean and variance of X can be used to assess the reliability of X. A reliability coefficient cannot be estimated because there is only one sample. Instead a CV, defined as the standard deviation of X divided by the mean of X x 100, is often used:

$$\text{CV \%} = \frac{\text{s.d.X}}{\bar{X}} \times 100.$$

A small CV is considered to indicate a reliable measure.

The CV provides only limited information about measurement error for two reasons. First, this type of reliability study only assesses the laboratory error and excludes errors due to storage and handling of specimens, and to the variation in the measure over time within individuals, which are usually greater sources of error in epidemiologic studies than the laboratory error. Second, the CV cannot be used to even assess the effect of laboratory error on the odds ratio. This is due to the fact that the CV is an estimate of the ratio of the standard deviation of X (which is an estimate of the standard deviation of the error (σ_E), as the true value of T is the same for each replicate) to $\bar{X}$, but it is ρ_{TX} which is a function of the ratio of the error variance to the total variance in X in the population of interest, that is needed to understand the impact of measurement error.

Reliability studies of binary biomarkers

Issues in the design and interpretation of reliability studies of binary biomarkers are similar to the issues discussed above for continuous biomarkers. However, the parameter used to measure the reliability of binary biomarkers is kappa (κ) rather than the intraclass correlation coefficient (44).

To compute κ for a binary marker, subjects are cross-classified by results on their first and second repeated measurements into a 2x2 table as follows:

		Measure 2		
		+	–	
Measure 1	+	p_{11}	p_{12}	r_1
	–	p_{21}	p_{22}	r_2
		s_1	s_2	1

where p_{11} is the proportion of reliability study subjects classified as positive on both measures, p_{12} is the proportion classified as positive on measure 1 but negative on measure 2, etc. Note that the four proportions (p_{ij}) sum to 1. The overall (marginal) proportions of those who are positive and negative for measure 1 are r_1 and r_2 respectively, and the marginal proportions on the second measure are s_1 and s_2.

One measure of agreement is the observed proportion for whom there was agreement. The observed proportion of agreement, P_o, is the sum of the proportions on the diagonal:

$$P_o = p_{11} + p_{22}.$$

However, this simple measure does not take into consideration the agreement that would be expected by chance. For example, suppose the first reader of a stain on a slide accurately classified 10% of subjects as positive and 90% as negative, but the second reader simply classified all slides as negative. Then the percentage agreement would be 90%, which does not reflect the poor repeatability across readers.

Kappa is a measure of agreement that corrects for the agreement expected by chance. The expected agreement by chance (on the diagonal), P_e, is:

$$P_e = r_1 s_1 + r_2 s_2.$$

Kappa is the observed agreement beyond chance divided by the maximum possible agreement beyond chance, and is estimated as:

$$\hat{k} = \frac{P_O - P_e}{1 - P_e}.$$

Kappa ranges from zero (no agreement beyond chance) to 1 (perfect agreement), although it can be less than zero if agreement is less than expected by chance. (See (12,45) for the computation of confidence intervals for κ.)

Similar to the concepts discussed above for continuous biomarkers, the results of a reliability study of a binary biomarker can, in some situations, be used to estimate the impact of biomarker error in the parent epidemiologic study that will use the biomarker. If the reliability study meets the assumptions of equal and uncorrelated error of the parallel test model (described above), and of nondifferential measurement error, then κ can be used to estimate the bias in the odds ratio. Specifically, it has been shown that under these assumptions, this equation provides an approximation of the attenuation of the odds ratio (46):

$$OR_o = OR_T^{\sqrt{\kappa}}. \qquad (10)$$

Similar to continuous measures, when the reliability study of a binary maker does not capture all sources of error (i.e. when some sources of error are repeated within-subjects (correlated)), κ will be artificially too high. Therefore, the attenuation of the odds ratio will be even greater than that predicted by Equation 10.

Review and conclusion

Before embarking on an epidemiologic study that uses a biomarker, it is important to research and understand the measurement error in that biomarker. This can be accomplished by reading previously published works on validity/reliability studies of the biomarker of interest, or conducting a new validity/reliability study. Measurement error in a biomarker refers to the error of a specific biomarker test, as applied in a specific way to a specific population, versus the true (etiologic) exposure. In epidemiologic studies, this error includes not only laboratory error, but also errors (variations) introduced during specimen collection, handling and storage, and due to month-to-month and year-to-year within-person variability of the biomarker.

Validity studies compare the biomarker to be used in an epidemiologic study to a perfect or near-perfect measure on a sample of subjects. The parameters used to quantify the error in a binary marker are sensitivity and specificity. For a continuous biomarker, X, the validity can be estimated by the bias ($\overline{X}$ - $\overline{T}$) and by the validity coefficient ρ_{XT} (correlation coefficient of X with T), where T is the (continuous) measure of the true exposure. To assess whether the error is differential between those with and without the disease, separate analyses on a group of cases and a group of controls are needed.

Often a perfect measure of the exposure is not available, so reliability (repeatability) studies are conducted. For these, a sample of subjects is measured twice using the same marker to measure errors (variations) in the biomarker over time, between laboratory technicians, etc. The reliability study should be designed to capture all sources of error, so that the error in one measure is not repeated in (correlated with) the errors in the other measures. To design a reliability study without correlated error, the repeated specimens for each person must be collected at different times over the relevant etiologic time period and handled, stored, and analysed with the degree of variation (different specimen collectors/laboratory technicians/batches) as would occur in the parent epidemiologic study. Reliability studies are analysed using κ for binary biomarkers and the intraclass correlation coefficient for continuous biomarkers.

Equations 1, 2, 6, 8 and 10 can be used to interpret these parameters from validity or well-designed reliability studies to estimate the degree of bias in the risk ratio in an epidemiologic study that will use the biomarker. These equations assume nondifferential measurement error (i.e. equal biomarker error for those with and without the disease). Nondifferential measurement error in the biomarker attenuates the risk ratio in an epidemiologic study towards the null value of one. This attenuation is often quite substantial, even for reasonably accurate biomarker measures. For continuous markers, the impact of nondifferential measurement error depends only on the validity coefficient ρ_{XT}, and not on the bias (Equation 6).

Differential biomarker error between those with the disease and those without can bias the risk ratio in any direction, and even make a risk factor appear to be protective. Thus, differential error can completely invalidate an epidemiologic study and must be avoided. Differential measurement error is a particular concern in case–control studies (and among the early cases in cohort studies) when the biomarker is not a fixed marker (e.g. genotype), and, therefore could be influenced by the disease, its preclinical phase, or its treatment. Assessment of differential error requires specimens on a sample of cases years before diagnosis and comparable early specimens on controls. This measure can serve as the “true” marker, even if it is not perfect, as long as it does not have differential error. For continuous variables, differential bias has the most problematic effects on the risk ratio; it could be estimated by comparing bias ($\overline{X}$ - $\overline{T}$) for cases with bias ($\overline{X}$ - $\overline{T}$) among controls.

One goal of giving examples of the large effects that even moderate degrees of biomarker measurement error have on epidemiologic studies, is to motivate attention to reducing the biomarker error. The researcher should focus on reducing the errors through appropriate quality control techniques for specimen collection, storage and laboratory analyses, and, if needed, by use of multiple measures over time in the parent epidemiologic study to reduce the errors caused by biomarker variation over time. These important methods and additional approaches to reduce biomarker error in epidemiologic studies are covered in other chapters of this book and by (12,15).

References

1. Walker AM, Blettner M (1985). Comparing imperfect measures of exposure. *Am J Epidemiol,* 121:783–790. PMID:4014171

2. de Klerk NH, English DR, Armstrong BK (1989). A review of the effects of random measurement error on relative risk estimates in epidemiological studies. *Int J Epidemiol,* 18:705–712.doi:10.1093/ije/18.3.705 PMID:2807678

3. Armstrong BG, Whittemore AS, Howe GR (1989). Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Stat Med,* 8:1151–1163, discussion 1165–1166.doi:10.1002/sim.4780080916 PMID:2799135

4. Dunn G. Design and analysis of reliability studies. London: Edward Arnold and New York: Oxford University Press; 1989.

5. White E. Effects of biomarker measurement error on epidemiological studies. In: Toniolo P, Boffetta P, Shuker DEG *et al.*, editors. Applications of biomarkers in cancer epidemiology. Lyon: IARC Scientific Publication; 1997. p. 73–94.

6. Chen TT (1989). A review of methods for misclassified categorical data in epidemiology. *Stat Med,* 8:1095–1106, discussion 1107–1108.doi:10.1002/sim.4780080908 PMID:2678350

7. Thomas D, Stram D, Dwyer J (1993). Exposure measurement error: influence on exposure-disease. Relationships and methods of correction. *Annu Rev Public Health*, 14:69–93.doi:10.1146/annurev.pu.14.050193.000441 PMID:8323607

8. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. 2nd ed. London/Boca Raton: Chapman and Hall CRC Press; 2006.

9. Holford TR, Stack C (1995). Study design for epidemiologic studies with measurement error. *Stat Methods Med Res,* 4:339–358. doi:10.1177/096228029500400405 PMID:8745130

10. Bashir SA, Duffy SW (1997). The correction of risk estimates for measurement error. *Ann Epidemiol,* 7:154–164.doi:10.1016/S1047-2797(96)00149-4 PMID:9099403

11. Thürigen D, Spiegelman D, Blettner M *et al.* (2000). Measurement error correction using validation data: a review of methods and their applicability in case-control studies. *Stat Methods Med Res,* 9:447–474.doi:10.1191/096228000701555253 PMID:1119 1260

12. White E, Armstrong BK, Saracci R. Principles of exposure measurement in epidemiology. 2nd ed. Collecting, evaluating and improving measures of disease risk factors. Oxford: Oxford University Press; 2008.

13. Vineis P. Sources of variation in biomarkers. In: Toniolo P, Boffetta P, Shuker DEG *et al*., editors. Applications of biomarkers in cancer epidemiology. Lyon: IARC Scientific Publication; 1997. p. 59–72.

14. Fleiss JL. The design and analysis of clinical experiments. New York (NY): John Wiley and Sons; 1986.

15. Tworoger SS, Hankinson SE (2006). Use of biomarkers in epidemiologic studies: minimizing the influence of measurement error in the study design and analysis. *Cancer Causes Control,* 17:889–899.doi:10.1007/s10552-006-0035-5 PMID:16841256

16. White E (2003). Design and interpretation of studies of differential exposure measurement error. *Am J Epidemiol,* 157:380–387.doi:10.1093/aje/kwf203 PMID:12615602

17. Kotz S, Johnson NL, editors. Encyclopedia of statistical sciences, vol 8. New York (NY): John Wiley and Sons; 1988.

18. Bross I (1954). Misclassification in 2 x 2 tables. *Biometrics*, 10:478–486 doi:10.2307/3001619.

19. Newell DJ (1962). Errors in the interpretation of errors in epidemiology. *Am J Public Health Nations Health,* 52:1925–1928. doi: 10.2105/AJPH.52.11.1925 PMID:13938241

20. Gullen WH, Bearman JE, Johnson EA (1968). Effects of misclassification in epidemiologic studies. *Public Health Rep,* 83:914–918. PMID:4972198

21. Goldberg JD (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *J Am Stat Assoc,* 70:561–567 doi:10.2307/2285933.

22. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH (1977). Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol,* 105:488–495. PMID:871121

23. Barron BA (1977). The effects of misclassification on the estimation of relative risk. *Biometrics*, 33:414–418.doi:10.2307/2529795 PMID:884199

24. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research. Belmont (CA): Lifetime Learning Publications; 1982. p. 183–193, 220–241.

25. Wu IC, Wu DC, Lu CY *et al.* (2004). Comparison of serum and urine ELISA methods for the diagnosis of Helicobacter pylori–a prospective pilot study. *Hepatogastroenterology*, 51:1736–1741. PMID:15532816

26. Hansen MH, Hurwitz WN, Bershad M (1961). Measurement errors in censuses and surveys. *Bull Int Stat Inst,* 38:359–374.

27. Lord FM, Novick MR. Statistical theories of mental test scores. Reading (MA): Addison-Wesley; 1968.

28. Cochran WG (1968). Errors of measurement in statistics. *Technometrics*, 10:637–666 doi:10.2307/1267450.

29. Nunnally JC. Psychometric theory. 2nd ed. New York (NY): McGraw-Hill; 1978. p. 190–225.

30. Allen MJ, Yen WM. Introduction to measurement theory. Monterey (CA): Brooks/Cole; 1979. p. 1–117.

31. Bohrnstedt GW. Measurement. In Rossi PH, Wright JD, Anderson AB, editors. Handbook of survey research. Orlando (FL): Academic Press; 1983. p. 70–121.

32. Fuller WA. Measurement error models. New York (NY): John Wiley and Sons; 1987.

33. Prentice RL (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69:331–342 doi:10.1093/biomet/69.2.331.

34. Whittemore AS, Grosser S. Regression methods for data with incomplete covariates. In: Moolgavkar SH, Prentice RL, editors. Modern statistical methods in chronic disease. New York (NY): John Wiley and Sons; 1986. p. 19–34.

35. Rosner B, Willett WC, Spiegelman D (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med,* 8:1051–1069, discussion 1071–1073.doi:10.1002/sim.4780080905 PMID:2799131

36. White E, Kushi LH, Pepe MS (1994). The effect of exposure variance and exposure measurement error on study sample size: implications for the design of epidemiologic studies. *J Clin Epidemiol*, 47:873–880.doi:10.1016/0895-4356(94)90190-2 PMID:7730890

37. Wu ML, Whittemore AS, Jung DL (1986). Errors in reported dietary intakes. I. Short-term recall. *Am J Epidemiol,* 124:826–835. PMID:3766514

38. McKeown-Eyssen GE, Tibshirani R (1994). Implications of measurement error in exposure for the sample sizes of case-control studies. *Am J Epidemiol,* 139:415–421. PMID: 8109576

39. Law MR, Thompson SG (1991). Low serum cholesterol and the risk of cancer: an analysis of the published prospective studies. *Cancer Causes Control,* 2:253–261. doi:10.1007/BF00052142 PMID:1831389

40. Winawer SJ, Flehinger BJ, Buchalter J *et al.* (1990). Declining serum cholesterol levels prior to diagnosis of colon cancer. A time-trend, case-control study. *JAMA*, 263:2083–2085.doi:10.1001/jama.263.15.2083 PMID:2319669

41. Gregorio DI, Marshall JR, Zielezny M (1985). Fluctuations in odds ratios due to variance differences in case-control studies. *Am J Epidemiol,* 121:767–774.doi:10.1093/aje/121.5.767 PMID:4014168

42. Missmer SA, Spiegelman D, Bertone-Johnson ER *et al.* (2006). Reproducibility of plasma steroid hormones, prolactin, and insulin-like growth factor levels among premenopausal women over a 2- to 3-year period. *Cancer Epidemiol Biomarkers Prev,* 15:972–978.doi:10.1158/1055-9965.EPI-05-0848 PMID:16702379

43. Garber CC, Carey RN. Laboratory statistics. In Kaplan L, Pesce A, editors. Clinical chemistry: theory, analysis, and correlation. St. Louis (MO): Mosby; 1984. p. 290–292.

44. Cohen J (1960). A coefficient of agreement for nominal scales. *Educ Psychol Meas,* 20:37–46 doi:10.1177/001316446002000104.

45. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. New York (NY): John Wiley and Sons; 2003.

46. Tavaré CJ, Sobel EL, Gilles FH (1995). Misclassification of a prognostic dichotomous variable: sample size and parameter estimate adjustment. *Stat Med,* 14:1307–1314.doi:10.1002/sim.4780141204 PMID:7569489