5

Statistical Methods for Assessing Biomarkers

Stephen W. Looney

1. Introduction

According to the *Dictionary of Epidemiology*, a *biomarker* is "a cellular or molecular indicator of exposure, health effects, or susceptibility" (***1***, p. 17). In this chapter, the primary focus is on markers of exposure, although the techniques described here can be applied to any type of biomarker. The process of assessing the quality of a biomarker consists of determining if the biomarker has adequate reliability and adequate validity. *Reliability* refers to "the degree to which the results obtained by a measurement procedure can be replicated" (***1***, p. 145). (Reliability is often used interchangeably with the terms *repeatability* and *reproducibility*.) The reliability of a measurement process is most often described in terms of intrarater and interrater reliability. *Intrarater reliability* (sometimes called *intraobserver agreement*) refers to the agreement between two different determinations made by the same individual and *interrater reliability* (sometimes called *interobserver agreement*) refers to the agreement between the determinations made by two different individuals. A reliable biomarker must exhibit adequate levels of both types of reliability. Also of concern in the assessment of the reliability of a biomarker are intersubject, intrasubject, and analytical measurement variability ***(2)***. The reliability of a biomarker must be established before validity can be examined; if the biomarker cannot be assumed to provide an equivalent result upon repeated determinations on the same biological material, it will not be useful for practical application.

The *validity* of a biomarker is defined to be the extent to which it measures what it is intended to measure. For example, Qiao et al. ***(3)*** proposed that the expression of a tumor-associated antigen by exfoliated sputum epithelial cells

From: *Methods in Molecular Biology, vol. 184: Biostatistical Methods*
Edited by: S. W. Looney © Humana Press Inc., Totowa, NJ

could be used as a biomarker in the detection of preclinical, localized lung cancer. For their biomarker to be valid, there must be close agreement between the classification of a patient (cancer/no cancer) using the biomarker and the diagnosis of lung cancer using the gold standard (in this case, consensus diagnosis using "best information"). As another example, body-fluid levels of cotinine have been proposed for use as biomarkers of environmental tobacco smoke exposure ***(4)***. For cotinine level to be a valid biomarker of tobacco exposure, it must be the case that high levels of cotinine correspond to high levels of tobacco exposure and low levels of cotinine correspond to low levels of exposure.

Both reliability and validity have to do with interchangeability. Adequate intrarater reliability means that there is minimal within-rater variability so that regardless of when the analyst performs the biomarker determination, we can safely assume that he or she will produce an equivalent result. Adequate interrater reliability means that there is minimal between-rater variability so that regardless of which analyst performs the biomarker determination, we can safely assume that equivalent results will obtain. Adequate validity means that the biomarker determination can be substituted for the gold standard result (assuming that there is a gold standard) or for the standard test result if there is no gold standard.

The appropriate statistical methods for assessing the reliability and validity of a biomarker depend upon the level of measurement of the biomarker. In this chapter, we offer separate recommendations for dichotomous and continuous biomarkers.

2. Dichotomous Biomarkers

2.1. Assessing Reliability of a Dichotomous Biomarker

The same statistical methodology is applied when examining the intra- and interrater reliability of a dichotomous biomarker. Both involve measuring the agreement between two different determinations of the biomarker status of an individual. To assess intrarater reliability, the same analyst would make the determination using the same specimen of material under "identical" conditions. This determination must be blinded, of course, so that the analyst is unaware on the second occasion that he or she is examining the same experimental material that he or she examined on the first occasion. To assess interrater reliability, two different analysts would make the determination using the same specimen of material under "identical" conditions. This determination should also be blinded so that Analyst A is unaware of the result of Analyst B and vice versa. For both intra- and interrater reliability, a 2×2 table is used to show the agreement (and disagreement) between the two determinations.

To assess intrarater reliability, the 2×2 table given in **Table 1** is constructed. To assess interrater reliability, a similar 2×2 table is constructed to show the agreement (and disagreement) between the two determinations made by different individuals on the same biological specimen.

Table 1
2 × 2 Table Showing Agreement Between Two Determinations by the Same Analyst of the Same Biological Specimen (Intrarater Reliability)

	Determination 2		
Determination 1	Positive	Negative	Total
Positive	a	b	f_1
Negative	c	d	f_2
Total	g_1	g_2	n

Table 2
2 × 2 Table Showing Agreement Between a Pathologist and a Cytotechnologist When Scoring the Same Stained Specimen

	Cytotechnologist		
Pathologist	Positive	Negative	Total
Positive	31	1	32
Negative	0	91	91
Total	31	92	123

Adapted, with permission, from Table 4 of Tockman et al. ***(5)***.

For example, Tockman et al. ***(5)*** examined the use of murine monoclonal antibodies to a glycolipid antigen of human lung cancer as a biomarker in the detection of early lung cancer. As part of their assessment of the interrater reliability of scoring stained specimens, they compared the results obtained on 123 slides read by both a pathologist and a cytotechnologist. They obtained the results given in **Table 2**.

Once the appropriate 2 × 2 table has been constructed, it is desirable to calculate a single numerical quantity as a measure of the reliability of the biomarker. The two most commonly used measures of agreement between two dichotomous variables are the *Index of Crude Agreement*, given by

$$p_0 = (a + d)/n, \tag{1.1}$$

and *Cohen's kappa*, given by

$$\kappa = (p_0 - p_e) / (1 - p_e),$$

where p_e = the percentage agreement between methods A and B that "can be attributed to chance" ***(6)***. The estimated percentage agreement between methods A and B that can be attributed to chance is given by

$$p_e = p_1 p_2 + q_1 q_2,$$

where $p_1 = (a+b)/n$, $p_2 = (a+c)/n$, $q_1 = 1-p_1$, *and* $q_2 = 1-p_2$. The formula for Cohen's kappa now becomes

$$\kappa = \frac{2(ad - bc)}{n^2 (p_1 q_2 + p_2 q_1)} \quad . \tag{1.2}$$

For the data given in **Table 2**, we obtain $\kappa = 0.979$ using **Eq. (1.2)**. This indicates excellent interrater reliability.

Kappa has the value 1 if there is perfect agreement ($b=c=0$), the value -1 if there is perfect disagreement ($a = d = 0$), and the value 0 if $p_0 = p_e$. Landis and Koch (***7***, p. 165) provide the following guidelines for interpreting the magnitude of kappa:

Value of κ	Interpretation
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

Cohen's kappa is the generally accepted method for assessing agreement between two dichotomous variables, neither of which can be assumed to be the gold standard ***(8)***, but several deficiencies have been noted [(***9***, p. 545), (***10***, p. 425)]. These deficiencies include: (1) If either method classifies no subjects into one of the two categories, $\kappa = 0$. (2) If there are no agreements for one of the two categories, $\kappa < 0$. (3) The value of κ is affected by the difference in the relative frequency of "disease" and "no disease" in the sample. The higher the discrepancy, the larger the value of p_e and the smaller the value of κ. (4) The value of κ is affected by any discrepancy between the relative frequency of "disease" for method A and the relative frequency of "disease" for method B. The greater the discrepancy, the smaller the expected agreement, and the larger the value of κ.

To adjust for these deficiencies, Byrt et al. ***(10)*** propose that, in addition to κ, one also report the prevalence-adjusted and bias-adjusted kappa (PABAK),

$$PABAK = \frac{(a+d) - (b+c)}{n} = 2p_0 - 1$$

where p_0 is the index of crude agreement given in **Eq. (1.1)**. (Note that PABAK is equivalent to the proportion of "agreements" between the two variables minus the proportion of "disagreements.")

As an illustration of some of the deficiencies of κ, consider the hypothetical data on the agreement between two observers given in **Table 3**.

Table 3
Hypothetical 2 × 2 Table Showing Agreement Between Two Observers

	Observer B		
Observer A	Positive	Negative	Total
Positive	80	15	95
Negative	5	0	5
Total	85	15	100

Even though the two observers agree on 80% of the specimens, the value of κ is –0.08, indicating poor agreement ***(7)***. Two of the previously mentioned deficiencies are at work here. First, because the two "observers" did not agree on any of the subjects who were classified as "negative," $\kappa < 0$. Second, the value of κ is adversely affected by the difference in the average relative frequencies of "disease" (90%) and "no disease" (10%) in the sample. The PABAK coefficient, which adjusts for both of these shortcomings, has the value $2p_0 - 1 = 2(0.80) - 1 = 0.60$. This is considered "moderate" agreement by the Landis and Koch criteria ***(7)*** and is a much more accurate representation than κ of the agreement between the two observers suggested by **Table 3**.

In additional to using κ and the PABAK coefficient to measure overall agreement, it is also advisable to describe the agreement separately in terms of those specimens that appear to be positive and those that appear to be negative. Using measures of positive agreement and negative agreement in assessing reliability is analogous to using sensitivity and specificity in assessing validity in the presence of a gold standard (*see* **Subheading 2.2.1.**). Such measures can be used to help diagnose the type(s) of disagreement that may be present.

Cicchetti and Feinstein ***(11)*** proposed indices of *average positive agreement* (p_{pos}) and *average negative agreement* (p_{neg}) for this purpose:

$$p_{pos} = \frac{a}{(f_1 + g_1)/2} \tag{1.3}$$

$$p_{neg} = \frac{d}{(f_2 + g_2)/2}.$$

Note that the denominators of p_{pos} and p_{neg} are the average number of subjects that the two methods classify as positive and negative, respectively. For the data in **Table 3**, $p_{pos} = 2(80)/(95 + 85) = 88.9\%$ and $p_{neg} = 2(0)/(5 + 15) = 0.0\%$. Thus, there is moderate overall agreement between the two observers (as measured by the PABAK coefficient of 0.60), "almost perfect agreement" on specimens

Table 4
"Truth Table" for Tumor-Associated Antigen as a Biomarker for Lung Cancer

	Gold standard		
Biomarker	Positive	Negative	Total
Positive	42	23	65
Negative	15	53	68
Total	57	76	133

Adapted, with permission, from Table 3 of Qiao et al. ***(3)***.

that appear to be positive, and no agreement on specimens that appear to be negative. Thus, efforts to improve the biomarker determination process should be targeted toward those specimens that are negative. (Several computationally intensive methods for estimating sensitivity and specificity in the absence of a gold standard have been proposed [e.g., ***12–14***]; however, these are beyond the scope of this chapter.)

2.2. Assessing Validity of a Dichotomous Biomarker

2.2.1. Gold Standard Is Available

Just as in the assessment of reliability described in the preceding, the assessment of the validity of a dichotomous biomarker involves the use of a 2 × 2 table. If a gold standard is available for the exposure or outcome that the biomarker is intended to represent (the "event"), then the term *conformity* is used to describe the agreement between the biomarker and the occurrence of the event and the term *truth table* is used to describe the 2 × 2 table.

For example, Qiao et al. ***(3)*** examined the agreement between the biomarker proposed by Tockman et al. ***(5)*** and the gold standard method for diagnosing lung cancer. The truth table for their data is given in **Table 4**.

The three measures of conformity obtained from this table are (1) *sensitivity* = a /(a + c) = 42/57 = 73.7%, the percentage of those that experienced the event that the biomarker correctly identified; (2) sp*ecificity* = d/(b + d) = 53/76 = 69.7%, the percentage of those that did not experience the event that the biomarker correctly identified; and ***(3)*** *accuracy* = (a + d)/n = (42 + 53)/133 = 71.4%, the percentage of all subjects that the biomarker correctly identified. Qiao et al. ***(3)*** compared these results with the "standard methods" of chest X-ray and sputum cytology and found that the biomarker proposed by Tochman et al. ***(5)*** had higher sensitivity, lower specificity, and slightly higher accuracy than both of the standard methods.

Table 5
Hypothetical 2 × 2 Table for Comparison of Two Biomarkers for Lung Cancer

	Sputum cytology		
Immunocytochemistry	Positive	Negative	Total
Positive	12	53	65
Negative	0	68	68
Total	12	121	133

2.2.2. Gold Standard Is Not Available

If a gold standard is not available for the exposure or outcome that the biomarker is intended to represent (the "event"), then the term *consistency* is used to describe the agreement between the biomarker and some other method used to determine if the event has occurred. This "other method" may be the "standard method" or a competing biomarker. The methods used for assessing intra- and interrater reliability described in **Subheading 2.1.** can be used to assess validity in this situation.

For example, suppose that no gold standard had been available in the study by Qiao et al. ***(3)*** referred to earlier. Then the investigators could have compared their biomarker based on immunocytochemistry with two "standard" methods of detecting preclinical, localized lung cancer (chest X-ray and sputum cytology). A hypothetical 2 × 2 table for the comparison of their biomarker with sputum cytology based on the assumption that their biomarker agreed with the sputum cytology result on all positive cases of the disease is given in **Table 5**.

Even though the two methods agree on 60% of the specimens, the value of κ is only 0.188, indicating slight agreement ***(7)***. Using the PABAK coefficient provides little improvement: PABAK=$2p_0$–1=2(0.602)–1=0.203. The indices of positive and negative agreement are $p_{pos} = 2(12)/(12 + 65) = 31.2\%$ and $p_{neg} = 2(68)/(121 + 68) = 72.0\%$, respectively. Thus, the disagreement between the two methods can be attributed primarily to those specimens that are thought to be positive. A similar analysis for the comparison of the biomarker based on immunocytochemistry with chest X-ray yields $\kappa = 0.483$ and PABAK = 0.489, indicating only moderate agreement (data not shown). The indices of positive and negative agreement are $p_{pos} = 64.6\%$ and $p_{neg} = 80.0\%$, indicating once again that the disagreement between the two methods can be attributed primarily to those specimens that are thought to be positive.

Table 6
Hypothetical Data on the Agreement Between Measurements A and B

Specimen number	Measurement A	Measurement B
1	31	206
2	4	28
3	17	112
4	14	98
5	16	104
6	7	47
7	11	73
8	4	43
9	14	93
10	7	57
11	10	87

3. Continuous Biomarkers

3.1. Use of Pearson's Correlation Coefficient

The most commonly used method for measuring agreement between two continuous variables X and Y is Pearson's correlation coefficient (PCC), denoted by r. However, at least as far back as 1973, it was recognized that the PCC is not appropriate for this purpose ***(15)***. The PCC measures *strength of linear association* between two variables, not agreement. We have perfect *agreement* between X and Y if and only if all points in a scatterplot of Y vs X lie along the line $Y = X$; however, we have perfect *correlation* between Y and X if all points in the scatterplot lie along *any* straight line. There are several other shortcomings of the PCC as a measurement of agreement; for example, it is dependent on the heterogeneity of the sample measurements and it is not related to the scale of measurement or to the size of error that might be clinically allowable (*see* [***15,16***] for further discussion).

To illustrate how the PCC can be misleading as a measure of agreement, consider the hypothetical data presented in **Table 6**. A useful first step in assessing *agreement* between X and Y is to construct the scatterplot and then superimpose the line $Y = X$ to get an idea of the deviation of the agreement between X and Y from 1.0. The data in **Table 6** are displayed in this manner in **Fig. 1**. The PCC between measurements A and B is almost perfect, $r = 0.989$, yet there is an obvious deviation from perfect agreement, with the value for measurement A being consistently less than the corresponding value for

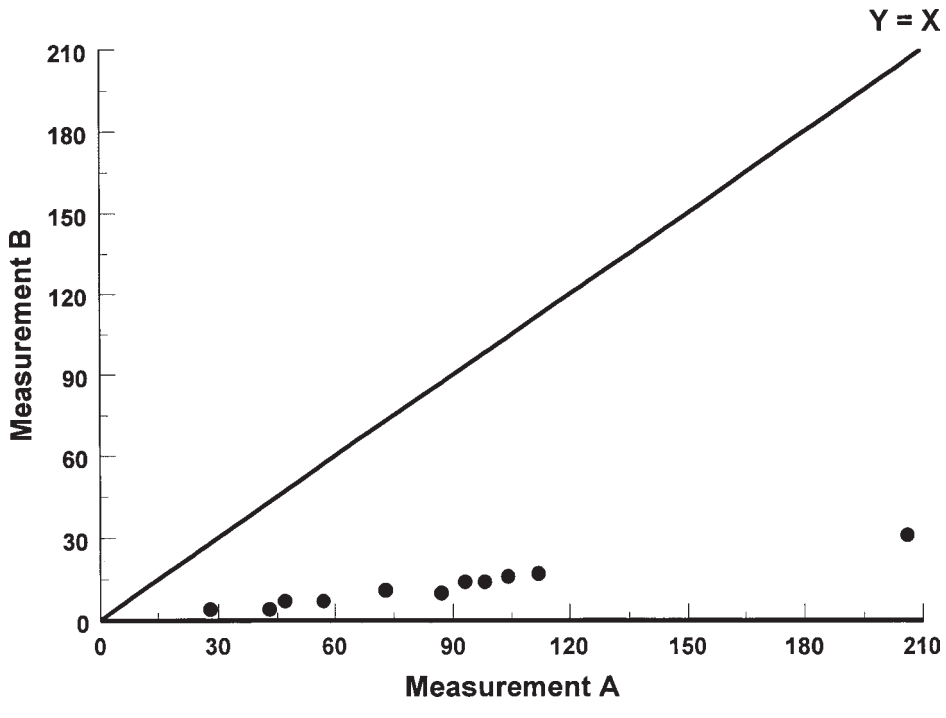

Fig. 1. Scatterplot of hypothetical data on agreement between biomarkers *A* and *B* with the line of perfect agreement ($Y = X$) superimposed.

measurement *B*. In the subheadings that follow, we describe alternatives to the PCC for measuring agreement and make recommendations for their appropriate use.

3.1.1. Assessing Intra- and Interrater Reliability

An alternative to the PCC that has been recommended for use in measuring agreement between two continuous measurements, neither of which is the gold standard, is the *intraclass correlation coefficient* (ICC), denoted by r_I ***(17)***. To assess both intra- and interrater reliability, a biomarker determination will be made repeatedly for each of *n* specimens. For intrarater reliability, the same specimen will typically be analyzed on two separate occasions by the same observer. For interrater reliability, the same specimen will typically be analyzed by two different observers. The ICC measures the size of the within-specimen variability relative to the between-specimen variability. It ranges between a value of 0, with $r_I = 0$ indicating no reproducibility at all (large within-speci-

men variability and zero between-specimen variability), and a value of 1, with $r_I = 1$ indicating perfect reproducibility (large between-specimen variability and zero within-specimen variability). Fleiss ***(18)*** provided guidelines for interpreting the magnitude of r_I:

Value of r_I	Interpretation
<0.40	Poor
0.40 – 0.75	Fair to good
0.75 – 1.00	Excellent

Suppose that n specimens are each repeatedly analyzed m times (replicates). (Typically, $m = 2$ for both intra- and interrater reliability.) We will assume that the n specimens constitute a random sample. We will also assume that the m replicates (the *occasions* at which the biomarker determinations are made in the case of intrarater reliability and the *observers* in the case of interrater reliability) also constitute a random sample. The simplest method for calculating the appropriate ICC under these assumptions is to use the two-way random effects model without interaction:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}; \qquad 1 \le i \le n; \qquad 1 \le j \le m$$

where

Y_{ij} = biomarker value for specimen i and replicate j
μ = population mean response
α_i = offset in mean response for specimen i
β_j = offset in mean response for replicate j
ε_{ij} = biomarker measurement error

The assumptions that underlie this model are as follows: $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\beta_j \sim N(0, \sigma_\beta^2)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, where $N(\theta, \delta^2)$ denotes the normal (Gaussian) distribution with mean θ and variance δ^2. All of these assumptions taken together imply that $Y_{ij} \sim N(\mu, \sigma^2)$, where $\sigma^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2$. The population value of the ICC is defined to be $\rho_I = \sigma_\alpha^2/\sigma^2$ and the sample value r_I is obtained by

$$r_I = \max\left[0, \widehat{\sigma}_\alpha^{\,2} / \widehat{\sigma}^{\,2}\right]$$

where $\widehat{\sigma}_\alpha^{\,2}$ and $\widehat{\sigma}^{\,2}$ are the sample estimates of the variance components from the two-way random effects model. After some simplification,

$$r_I = \frac{MSS - MSE}{MSS + (m-1)MSE + m(MSR - MSE)/n} \tag{3.1}$$

where *MSE* denotes the mean square due to error, *MSS* denotes the mean square due to specimens, *MSR* denotes the mean square due to replicates, m denotes the number of replicates for each specimen, and n denotes the number of specimens. The formulas for calculating these mean squares are as follows:

$$MSS = m\sum_{i=1}^{n}\left(\bar{y}_{i.} - \bar{\bar{y}}\right)^2/(n-1)\ ,$$

$$MSR = n\sum_{j=1}^{m}\left(\bar{y}_{.j} - \bar{\bar{y}}\right)^2/(m-1)\ , \qquad (3.2)$$

$$MSE = \left(\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij} - \bar{\bar{y}}\right)^2 - (n-1)\,MSB - (m-1)\,MSR\right)/(nm - n - m + 1)$$

where

$$\bar{y}_{i.} = \sum_{j=1}^{m} y_{ij}/m,\ \bar{y}_{.j} = \sum_{i=1}^{n} y_{ij}/n,\ \bar{\bar{y}} = \sum_{i=1}^{n} \bar{y}_{i.}/n\ .$$

As an example, consider the data in **Table 7,** taken from a study of a bile-acid- induced apoptosis assay for colon cancer risk ***(19)***. This is an example of evaluating interrater reliability with $n = 15$ and $m = 2$. Applying the formulas in **Eqs. (3.1)** and **(3.2)**, we obtain $MSS = 698.919$, $MSR = 246.533$, and $MSE = 43.176$ and

$$r_I = \frac{698.919 - 43.176}{698.919 = (2-1)\,(43.176) + 2(246.533 - 43.176)\,/15} = 0.8525\ ,$$

which indicates excellent interrater reliability ***(18)***.

An alternative to the ICC that is useful in evaluating the intra- and interrater reliability of biomarkers is *Lin's coefficient of concordance* ***(20)***, defined in the population to be

$$\rho_c = 1 - \frac{E\left[(X_1 - X_2)^2\right]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}\ ,$$

where

μ_1 = mean of X_1
μ_2 = mean of X_2
σ_1^2 = variance of X_1
σ_2^2 = variance of X_2

or, in the case $\sigma_1^2 = \sigma_2^2$

$$\rho_c = \frac{\rho}{1 + \left(\dfrac{\mu_2 - \mu_2}{\sigma\sqrt{2}}\right)^2}\ .$$

The corresponding sample quantity, r_c , is

$$r_c = \frac{2s_{12}}{s_1^2 + s_2^2 + (\bar{x}_1 - \bar{x}_2)^2}$$

Table 7
Interrater Reliability Data from a Study of a Bile-Induced Apoptosis Assay for Colon Cancer Risk *(19)*

Specimen	Observer 1	Observer 2
1	11	27
2	9	15
3	54	72
4	55	63
5	50	65
6	44	49
7	58	51
8	5	8
9	21	30
10	58	43
11	41	40
12	59	62
13	39	52
14	34	49
15	23	21

Data courtesy of Carol Bernstein, personal communication, October 17, 2000.

where

s_{12} = sample covariance of X_1 and X_2

$\bar{x}_1$ = sample mean of X_1

$\bar{x}_2$ = sample mean of X_2

s_1^2 = sample variance of X_1

s_2^2 = sample variance of X_2.

It can be shown that $r_c = 1$ if there is perfect agreement between the sample values of X_1 and X_2, $r_c = -1$ if there is perfect negative agreement, and $-1 < r_c < 1$ otherwise. The interpretation of the value of Lin's coefficient is the same as that for the ICC given earlier. The calculation of r_c for the data given in **Table 7** proceeds as follows:

$$\bar{x}_1 = 37.40,\ \bar{x}_2 = 43.13,\ s_1^2 = 368.543,\ s_2^2 = 373.552,\ s_{12} = 327.9990\ .$$

Therefore,

$$r_c = \frac{2s_{12}}{s_1^2 + s_2^2 + (\bar{x}_2 - \bar{x}_2)^2} = \frac{2(327.999)}{368.544 + 373.5523 + (37.4 - 43.13)^2} = 8.847\ ,$$

an almost identical result to the ICC of 0.853. The PCC for these data is 0.884.

For an example of data that exhibit strong correlation, but poor agreement, again consider the data in **Table 6**. The PCC between *X* and *Y* is almost perfect, $r = 0.989$; however, the intraclass correlation is zero, and Lin's coefficient is only 0.102. The PCC indicates near-perfect linear association, but both of the latter coefficients indicate extremely poor agreement, a much more accurate representation of what is indicated by the plot in **Fig. 1**. In **Subheading 3.2.2.2.**, we present a method for the detailed analysis of the disagreement between two measurements.

3.1.2. Assessing Intersubject, Intrasubject, and Analytical Measurement Variability

The approach described in this subheading very closely follows the scheme proposed by Taioli et al. ***(2)*** for evaluating the reliability of a biomarker. In addition to intra- and interrater reliability, there are three major components of biomarker variability that must be considered when evaluating reliability. These are: *intersubject* variability, sources of which might include genetics, race, gender, diet; *intrasubject* variability, sources of which include random biologic variation, change in diet, change in exposure; and *analytical* or *laboratory* variability, sources of which include variation between analytical batches, variation within analytical batches, and random variation within the measurement process itself. As Taioli et al. (***2***, p. 308) point out, even if a biomarker has acceptable validity, an excess of intraindividual and/or laboratory variability might render it unusable for research purposes.

To examine each of the sources of variability mentioned previously, biological specimens from each of *n* subjects are analyzed on *m* occasions (e.g., *weeks*), and the biomarker determination is repeated for each of *r* replicate samples (e.g., *aliquots*) from each specimen on each occasion. Replicate samples must be used for all aspects of a proper assessment of biomarker reliability to be examined.

The data from these biomarker determinations are used to estimate the components of biomarker variability mentioned previously. The data from multiple subjects are used to assess intersubject variability, data from multiple occasions are used to assess intrasubject variability, and data from replicate samples are used to assess analytical variability. Once an estimate of analytical variability (or "error variance") is available, it can be used in method comparison studies (*see* **Subheading 3.2.2.2.**). The estimates of intersubject variability, intrasubject variability, and analytical variability can also be combined to form an estimate of the total variance of the biomarker determination, which is useful in calculating the appropriate sample size for future studies in which the biomarker will be used ***(2)***.

The statistical model that underlies the approach of Taioli et al. ***(2)*** is very similar to the one used in assessing inter- and intrarater reliability for continuous biomarkers (*see* **Subheading 3.1.1.**), except that now the replicate observations must be accounted for:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}\,;\ 1 \le i \le n;\ 1 \le j \le m;\ 1 \le k \le r \qquad (3.3)$$

where

Y_{ijk} = biomarker value for subject i on occasion j and replicate k
μ = population mean response
α_i = offset in mean response for subject i
β_j = offset in mean response for occasion j
$(\alpha\beta)_{ij}$ = offset in mean response for the interaction between subject i and occasion j
ε_{ijk} = biomarker measurement error

(A nonzero interaction term indicates that the differences among subjects vary from occasion to occasion.)

The assumptions that underlie model (3.3) are as follows:

$$\alpha_i \sim N(0,\sigma_\alpha^2),\ \beta_i \sim N(0,\sigma_\beta^2),\ (\alpha\beta)_{ij} \sim N(0,\sigma_{\alpha\beta}^2),\ \text{and}\ \varepsilon_{ijk} \sim N(0,\sigma_\varepsilon^2)\ .$$

All of these assumptions taken together imply that

$$Y_{ijk} \sim N(\mu,\sigma^2)\ ,\ \text{where}\ \sigma^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma_\varepsilon^2\ .$$

As in **Subheading 3.1.1.**, the appropriate statistical method for analyzing biomarker data of this type is *two-way random-effects analysis of variance (ANOVA).* This analysis will yield tests of significance for each main effect (subjects and occasions) and a test of the interaction between subjects and occasions. It will also provide estimates of each variance component (intersubject variability, intrasubject variability, and analytical variability).

Taioli et al. ***(2)*** provide an example of the application of their approach to assessing the reliability of four different biomarkers for exposure to carcinogenic metals. The biomarkers they examined are (1) DNA–protein crosslink (DNA–PC), (2) DNA–amino acid crosslink (DNA–AA), (3) metallothionein gene expression (MT), and (4) autoantibodies to oxidized DNA bases (DNAox). We consider the results of only one of their studies here (DNA–PC). In this study, weekly blood samples were drawn three times ($m = 3$) from each of five healthy, unexposed subjects ($n = 5$) and each blood sample was divided into either three or four aliquots ($r = 3$ or 4) for analysis. The blood samples were analyzed during the week in which they were drawn. The results of the random-effects ANOVA are given in **Table 8**. The error variance for the DNA–PC determination is estimated to be 0.0317 and the estimated total variance is $(0.0545 + 0.0176 + 0.0110 + 0.0317) = 0.1148$.

Table 8
Random-Effects ANOVA for DNA–Protein Cross-link Data

Variance component	Variance estimate	*F* (d.f.)	*p*-value
Week	0.0545	13.68 (2,7)	< 0.010
Between subject	0.0176	3.45 (4,7)	0.073
Week x subjects	0.0110	2.33 (7,40)	0.045
Error	0.0317	—	—

Adapted, with permission, from Table 2 of Taioli et al. ***(2)***.

From **Table 8**, one can see that there is a significant "week" effect (i.e., intrasubject variability), and that the "subject" effect (i.e., intersubject variability) does not quite reach statistical significance. There is also a significant interaction between "week" and "subjects"; this suggests that the "week" effect varies across subjects. The authors point out that, by analyzing the blood samples in the week in which they were drawn, they introduced a possible batch effect that is confounded with the "week" effect. Therefore, the significant intrasubject variability could be a result of the batch effect and not of true week-to-week variation. To prevent this batch effect in the future, the authors modified their assay so that the DNA–PC determination could be performed for all samples at one time.

3.2. Assessment of the Validity of a Continuous Biomarker

3.2.1. Gold Standard Is Available

The assessment of the validity of a continuous biomarker in the presence of a gold standard is equivalent to the calibration of the biomarker ***(21)*** and is beyond the scope of this chapter. Numerous detailed accounts of methods for calibrating a biomarker are already available (e.g., ***[22]***).

3.2.2. Gold Standard Is Not Available

This is equivalent to what is commonly referred to as a "method comparison study" ***(15,16,23)***. We have already noted the problems with using the PCC for measuring agreement between continuous variables and Westgard and Hunt (***15***, p. 53) go so far as to state that "the correlation coefficient … is of no practical use in the statistical analysis of comparison data." The ICC, which has been proposed as an alternative to the PCC for measuring agreement between two continuous variables ***(24)***, is also not appropriate as a measure of consistency between two different biomarkers, primarily because to use the version of the ICC recommended in ***(24)*** requires the assumption that the two biomarkers being considered are a random sample from the population of all

biomarkers (***25***, p. 338). There are other disadvantages of using the ICC for measuring agreement between two biomarkers, including some of the same disadvantages involved in using the PCC namely, that it is dependent on the heterogeneity of the sample measurements, and it is not related to the scale of measurement or to the size of error that might be clinically allowable ***(25,26)***. (Lin's coefficient of concordance is also affected by the heterogeneity of the sample ***[27]***, but see also ***[28]***.) In the next three subheadings, we present alternative methods that have none of the disadvantages of the ICC.

3.2.2.1. The Bland–Altman Method

An alternative method for measuring consistency between two biomarkers X_1 and X_2 in which both biomarker determinations are in the same units is to apply the methodology proposed by Altman and Bland ***(16,21)***. The steps involved in this approach are as follows:

1. Construct a scatterplot and superimpose the line $X_2 = X_1$.
2. Plot the difference between X_1 and X_2 (denoted by d) vs the mean of X_1 and X_2 for each subject.
3. Perform a visual check to make sure that the within-subject repeatability is not associated with the size of the measurement, that is, that the bias (as measured by $[X_1 - X_2]$) does not increase (or decrease) systematically as $(X_1 + X_2)/2$ increases.
4. Perform a formal test to confirm the visual check in **step 3** by testing the hypothesis H_0: $\rho = 0$, where ρ = the true correlation between $(X_1 - X_2)$ and $(X_1 + X_2)/2$.
5. If there is no association between the size of the measurement and the bias, then proceed to **step 6** below. If there does appear to be significant association, then an attempt should be made to find a transformation of X_1, X_2, or both so that the transformed data do not exhibit any association. This can be accomplished by repeating **steps 2–4** for the transformed data. The logarithmic transformation has been found to be most useful for this purpose. (If no transformation can be found, Altman and Bland ***[16]*** recommend describing the differences between the methods by regressing $[X_1 - X_2]$ on $[X_1 + X_2]/2$.)
6. Calculate the "limits of agreement"; $\bar{d} - 2s_d$ to $\bar{d} + 2s_d$, where $\bar{d}$ is the mean difference between X_1 and X_2 and s_d is the standard deviation of the differences.
7. Approximately 95% of the differences should fall within the limits in **step 6** (assuming a normal distribution). If the differences within these limits are not clinically relevant, then the two methods can be used interchangeably. However, it is important to note that this method is applicable *only* if both measurements are made in the same units.

For example, Bartczak et al. ***(29)*** compared a high-pressure liquid chromatography (HPLC)-based assay and a gas chromatography (GC)-based assay for urinary muconic acid, both of which have been used as biomarkers of exposure to benzene. Their data, after omitting an outlier due to an unresolved chromatogram peak, are given in **Table 9**.

Table 9
Data on Comparison of Determinations of Muconic Acid in Human Urine by HPLC–Diode Array and GC–MS Analysis

Specimen number	HPLC (X_1)	GC–MS (X_2)	$X_1 - X_2$	$(X_1 + X_2)/2$
1	139	151	–12.00	145.00
2	120	93	27.00	106.50
3	143	145	–2.00	144.00
4	496	443	53.00	469.50
5	149	153	–4.00	151.00
6	52	58	–6.00	55.00
7	184	239	–55.00	211.50
8	190	256	–66.00	223.00
9	32	69	–37.00	50.50
10	312	321	–9.00	316.50
11	19	8	11.00	13.50
12	321	364	–43.00	342.50

Adapted, with permission, from Table 2 of Bartczak et al. ***(29)***.

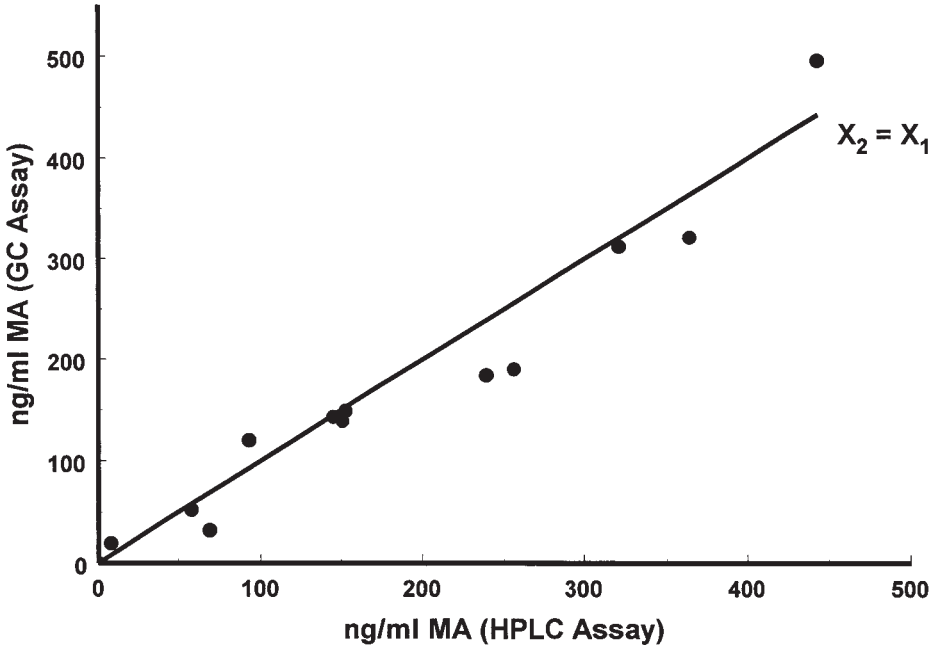

Fig. 2. Scatterplot of data on agreement between (HPLC)-based assay and (GC)-based assay for urinary muconic acid with the line of perfect agreement ($X_2 = X_1$) superimposed.

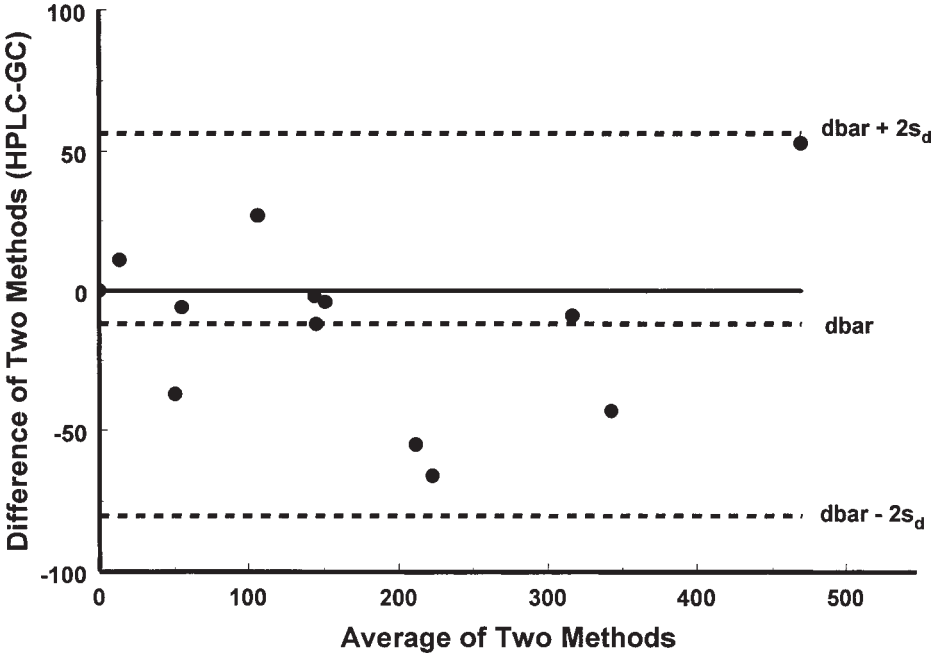

Fig. 3. Plot of difference vs mean for data on agreement between (HPLC)-based assay and (GC)-based assay for urinary muconic acid.

Figure 2 shows the scatterplot of X_2 vs X_1 with the line $X_2 = X_1$ superimposed. This plot indicates fairly good agreement except that 9 of the 12 data points are below the line of agreement. Figure 3 shows the plot of the difference (HPLC – GC) vs the mean of HPLC and GC for each subject. A visual inspection of Fig. 3 suggests that the within-subject repeatability is not associated with the size of the measurement, that is, that (HPLC – GC) does not increase (or decrease) systematically as (HPLC + GC)/2 increases. The sample correlation between (HPLC – GC) and (HPLC + GC)/2 is r = 0.113 and the p-value for the test of H_0: r = 0 is 0.728. Therefore, the assumption of the independence between the difference and the average is not contradicted by the data. The "limits of agreement" are $\bar{d} - 2s_d = -11.9 - 2(34.2) = -80.3$ to $\bar{d} + 2s_d = -11.9 + 2(34.2) = 56.5$ and these are represented (along with $\bar{d}$) by dotted lines in Fig. 3. (Note that all of the differences fall within the limits $\bar{d} - 2s_d$ to $\bar{d} + 2s_d$.) If differences as large as 80.3 are not clinically relevant, then the two methods can be used interchangeably. Given the order of magnitude of the measurements in Table 9, it would appear that a difference of 80 would be clinically important, so there appears to be inadequate agreement between the two methods. This was not obvious from the plot in Fig. 2.

3.2.2.2. Deming Regression

Strike ***(23)*** describes an approach for determining the type of disagreement that may be present when comparing two biomarkers. These methods are most likely to be applicable when one of the methods (method *X*) is a *reference* method, perhaps a biomarker that is already in routine use, and the other method (method *Y*) is a *test* method, usually a new biomarker that is being evaluated. Any systematic difference (or *bias*) between the two biomarkers is relative in nature, as neither method can be thought of as representing the true exposure.

As in the Bland–Altman method described in **Subheading 3.2.2.1.**, the first step is to construct a scatterplot of *Y* vs *X* and superimpose the line $Y = X$. Any systematic discrepancy between the two biomarkers will be represented on this plot by a general shift in the location of the points away from the line $Y = X$. Strike assumes that systematic differences between the two biomarkers can be attributed to either *constant bias*, *proportional bias*, or both, and assumes the following models for each biomarker result:

$$\begin{aligned} X_i &= \xi_i + \delta_i,\ 1 \le i \le n \\ Y_i &= \eta_i + \varepsilon_i,\ 1 \le i \le n \end{aligned} \tag{3.4}$$

where
X_i = observed value for biomarker *X*,
ξ_i = true value of biomarker *X*,
δ_i = random error for biomarker *X*,
Y_i = observed value for biomarker *Y*,
η_i = true value of biomarker *Y*,
ε_i = random error for biomarker *Y*.

Strike further assumes that the errors δ_i and ε_i are stochastically independent of each other and normally distributed with constant variance (σ_δ^2 and σ_ε^2, respectively) throughout the range of biomarker determinations in the study sample. (Strike points out that constant variance assumptions are usually unrealistic in practice and recommends a computationally intensive method for accounting for this lack of homogeneity. This method is incorporated into the MINISNAP software provided with Strike ***[23]***).

Strike assumes that any systematic discrepancy between methods *X* and *Y* can be represented by

$$\eta_i = \beta_0 + \beta_1 \xi_i \tag{3.5}$$

In this model, *constant bias* is represented by deviations of β_0 from 0 and *proportional bias* by deviations of β_1 from 1. (This is the same terminology used by Westgard and Hunt ***[15]***). If we now incorporate **Eq. (3.5)** into the equation for Y_i in **Eq. (3.4)**, we have

$$Y_i = \beta_0 + \beta_1 X_i + (\varepsilon_i - \beta_1 \delta_i). \tag{3.6}$$

Model (3.6) is sometimes called a *functional errors-in-variables model* and assessing agreement between biomarkers X and Y requires the estimation of the parameters β_0 and β_1. Strike proposes a method that requires an estimate of the ratio of the error variances given by $\lambda = \sigma_\varepsilon^2/\sigma_\delta^2$. This method is generally referred to in the clinical laboratory literature as "Deming regression"; however, this is somewhat of a misnomer as Deming was concerned with generalizing the errors-in-variables model to nonlinear relationships. Strike points out that the method he advocates for obtaining estimates of β_0 and β_1 is actually due to Kummel ***(30)***.

The equations for estimating β_0 and β_1 are as follows:

$$\hat{\beta}_1 = \frac{\left(S_{yy} - \hat{\lambda}S_{xx}\right) + \sqrt{\left(S_{yy} - \hat{\lambda}S_{xx}\right)^2 + 4\hat{\lambda}S_{xy}^2}}{2S_{xy}}, \tag{3.7}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X},$$

$$\hat{\lambda} = \hat{\sigma}_\varepsilon^2 / \hat{\sigma}_\delta^2,$$

where

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2,\ S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,\ S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

The estimate $\hat{\lambda}$ can be obtained either from error variance estimates for each biomarker provided by the laboratory or by estimating each error variance using

$$\hat{\sigma}^2 = \sum_{i=1}^{n} d_i^2 / (2n)$$

where d_i = difference between the two determinations of the biomarker (replicates) for specimen i. (The error variance can also be estimated from the assessment of reliability recommended by Taioli et al. ***[2]*** that is described in **Subheading 3.1.2.**) The methodology proposed by Strike cannot be applied without an estimate of the ratio of error variances of the two biomarkers.

To perform significance tests for β_0 and β_1, we need formulas for the standard errors (*SE*s) of $\hat{\beta}_0$ and $\hat{\beta}_1$. The approximations that Strike recommends for routine use are given by

$$SE(\hat{\beta}_1) = \left\{\frac{\hat{\beta}_1^2\left[(1 - r^2)/r^2\right]}{n - 2}\right\}^{1/2}$$

$$SE(\hat{\beta}_0) = \left\{\frac{\left[SE\left(\hat{\beta}_1\right)\right]^2 \sum X^2}{n}\right\}^{1/2} \tag{3.8}$$

where $$r^2 = [S_{xy}/(S_{xx}S_{yy})^{1/2}]^2$$

is the usual "R^2" value for the regression of Y on X. Tests of $H_0: \beta_1 = 1$ and $H_0: \beta_0 = 0$ can be performed by referring $(\hat{\beta}_1 - 1)/SE(\hat{\beta}_1)$ and $(\hat{\beta}_0)/SE(\hat{\beta}_0)$, respectively, to the $t(n-2)$ distribution.

As mentioned earlier, the approach described previously is based on the assumption that the error variances σ_δ^2 and σ_ε^2 are constant throughout the range of biomarker determinations in the study sample. However, as Strike points out, this assumption is usually unrealistic in practice and recommends the "weighted Deming regression" methods of Linnet ***(31,32)*** for accounting for this lack of homogeneity. These methods are incorporated into the MINISNAP software provided with Strike ***(23)***; however, replicate measurements are required for each test specimen using both biomarkers in order to apply these methods.

For example, consider the data in **Table 6** that were discussed in **Subheading 3.1.** The scatterplot of Y vs X in **Fig. 1** indicated substantial lack of agreement between X and Y and this was borne out by the intraclass correlation coefficient and Lin's coefficient, both of which indicated substantial disagreement. We can apply Strike's method to gain a better understanding of this disagreement.

Using the formulas in **Eqs. (3.7)** and **(3.8)**, we obtain $\hat{\beta}_1 = 0.158$, $SE(\hat{\beta}_1) = 0.007$, $\hat{\beta}_0 = -1.342$, $SE(\hat{\beta}_0) = 0.614$. For the test of $H_0: \beta_1 = 1$, this yields

$$t_{cal} = (\hat{\beta}_1 - 1)/SE(\hat{\beta}_1) = (0.158 - 1)/0.007 = -129.54,$$

and using a t-distribution with $n - 2 = 9$ degrees of freedom we find $p < 0.0001$. Therefore, there is significant proportional bias (which in this case is negative since $\hat{\beta}_1 < 1.0$). For the test of $H_0: \beta_0 = 0$, we have

$$t_{cal} = \hat{\beta}_0/SE(\hat{\beta}_0) = -1.342/0.614 = -2.19,$$

and, again using a t-distribution with 9 degrees of freedom, we have $p = 0.056$. Thus, the constant bias is not statistically significant, but just misses the usual cutoff of 0.05.

3.2.2.3. Spearman Correlation

A method that can be used to measure consistency between measurements that are in different units is *Spearman's rank correlation coefficient (SCC)*, denoted by r_s. This method is useful, and to be preferred over Pearson's correlation, when examining the agreement between two biomarkers whose determinations are in different units, or between a biomarker and some other

measure of exposure such as environmental monitoring. For example, Coultas et al. ***(33)*** used the SCC to measure the consistency between various measures of exposure to environmental tobacco smoke at work (e.g., nicotine exposure measured with a personal monitoring pump vs. post-shift urinary cotinine).

The SCC measures the agreement between two sets of measurements, after the measurements have been ordered ("ranked") from smallest to largest. Like other correlation coefficients, the SCC ranges between 1 (perfect agreement) and –1 (perfect negative agreement). As an example, suppose $n = 9$ and that the values obtained from Biomarker A for the nine specimens have been arranged in order from smallest to largest, with the smallest receiving rank 1 and the largest receiving rank 9. The same ordering is repeated for the nine specimens for each of biomarkers *B* and *C*. The specimens are arranged in order of their rankings according to biomarker *A* and then the ranks for biomarkers *B* and *C* are also noted:

Rank by biomarker *A*	1 2 3 4 5 6 7 8 9
Rank by biomarker *B*	1 2 3 4 5 6 7 8 9
Rank by biomarker *C*	9 8 7 6 5 4 3 2 1

According to the SCC, biomarkers *A* and *B* have perfect agreement ($r_s = 1$), whereas biomarkers *A* and *C* have perfect negative agreement ($r_s = -1$).

If r_s is close to 1, then we can assume that a subject with high levels of exposure, according to biomarker *A*, will also tend to have high levels of exposure, according to biomarker *B* (and similarly for low levels). Therefore, regardless of which biomarker we use, we can feel confident that subjects will be assigned to a high (or low) exposure group in a consistent manner. However, the PCC is not recommended for this purpose because *r* may be low even though high levels of exposure, according to biomarker *A*, are associated with high levels of exposure, according to biomarker *B*. This can occur, for example, if the relationship between the two biomarkers is nonlinear. Spearman's correlation will have a large value if the two biomarker determinations are strongly related according to *any* monotonic relationship.

The SCC is calculated using the following formula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} (r_i - s_i)^2}{n(n^2 - 1)}, \tag{3.9}$$

where r_i = the rank of subject *i* according to biomarker *A*, s_i = the rank of subject *i* according to biomarker *B*, and *n* = the number of subjects. Morton et al. ***(34)*** provide guidelines that can be used to interpret the value of r_s:

Table 10
Data on Concentrations of *o*-Cresol and Hippuric Acid Concentrations in Urine Samples

Specimen number	*o*-Cresol (μg/mL)	Rank of *o*-cresol	Hippuric acid (mg/mL)	Rank of Hippuric acid
1	0.21	1.5	0.30	2.0
2	0.21	1.5	0.80	5.0
3	0.25	3.0	0.40	3.0
4	0.28	4.0	0.50	4.0
5	0.32	5.0	1.10	7.5
6	0.34	6.0	1.19	9.0
7	0.41	7.0	1.30	12.0
8	0.44	8.5	1.08	6.0
9	0.44	8.5	1.10	7.5
10	0.51	10.0	1.20	10.5
11	0.59	11.0	1.20	10.5
12	0.76	12.0	1.33	13.0
13	1.25	13.0	0.20	1.0
14	1.36	14.0	2.10	14.0
15	2.80	15.0	3.02	15.0

Adapted, with permission, from Tables 1–3 of Amorim and Alvarez-Leite ***(35)***.

Value of $\lvert r_s \rvert$	Interpretation
0.00 – 0.20	Negligible
0.21 – 0.50	Weak
0.51 – 0.80	Moderate
0.81 – 1.00	Strong

To illustrate the calculation of the SCC, consider the data in **Table 10** on concentrations of *ortho*-cresol and hippuric acid in urine samples of workers exposed to toluene ***(35)***, which have been ranked according to the magnitude of the *o*-cresol value. Applying the formula in Eq. (3.9), we obtain $r_s = 0.632$, which indicates a moderate degree of consistency between the two measurements.

3.2.2.4. Criterion and Construct Validity

There are two types of validity that should be examined when evaluating a biomarker in the absence of a gold standard. *Criterion validity* is examined by correlating the biomarker with measures of some other phenomenon that is

expected to be correlated with the exposure or outcome that the biomarker represents. There are two types of criterion validity, *concurrent* and *predictive*. *Concurrent* refers to other phenomena that are contemporaneous with the biomarker, whereas *predictive* refers to phenomena that occur at some future time point.

For example, to assess concurrent validity in a study of the usefulness of post-shift urinary and salivary cotinine as a biomarker for workplace exposure to environmental tobacco smoke, urinary and salivary cotinine levels of nonsmoking workers were correlated with the total number of smokers and the total number of hours exposed to cigarette smoke in the workplace ***(33)***. As an example of predictive validity, in a study of the usefulness of plasma cotinine as a biomarker for environmental tobacco smoke, the authors examined the correlation between plasma cotinine and the metabolic clearance of theophylline, a drug whose metabolism is known to be increased in nonsmokers by the presence of cigarette smoke ***(36)***. The PCC is typically used to measure criterion validity; however, we recommend that the SCC be used instead, as the PCC measures only the degree of linear relationship, whereas the SCC is sensitive to any monotonic relationship between the biomarker and the criterion.

The other type of validity that should be evaluated in the absence of a gold standard is *construct validity*, which is examined in light of hypotheses formulated by the investigator about the characteristics of those who should have high levels of the exposure represented by the biomarker vs those who should have low levels. For example, Hüttner et al. ***(37)*** evaluated chromosomal aberrations in human peripheral blood lymphocytes as a biomarker of chronic exposure to heavy metals and dioxins/furans over a long period of time. As part of their examination of construct validity, they compared 52 exposed individuals from a polluted area with 51 matched controls from a distant nonindustrialized area and found a statistically significant increase in the frequency of chromosomal aberrations in human peripheral blood lymphocytes in the exposed group ($p < 0.001$). Construct validity is generally assessed by performing the appropriate statistical test to carry out the comparison of interest. For example, Hüttner et al. ***(37)*** used Fisher's exact test to compare the exposed and unexposed individuals in terms of the dichotomous outcome (chromosomal aberration/no chromosomal aberration). For continuous outcomes, the appropriate normal-theory test should be used if the outcome appears to follow a normal distribution (*t*-test for the comparison of two groups, one-way ANOVA for more than two groups). If the outcome data are highly skewed or otherwise non-normal, the Mann–Whitney–Wilcoxon test should be used to compare two groups, and the Kruskal–Wallis test should be used for more than two groups.

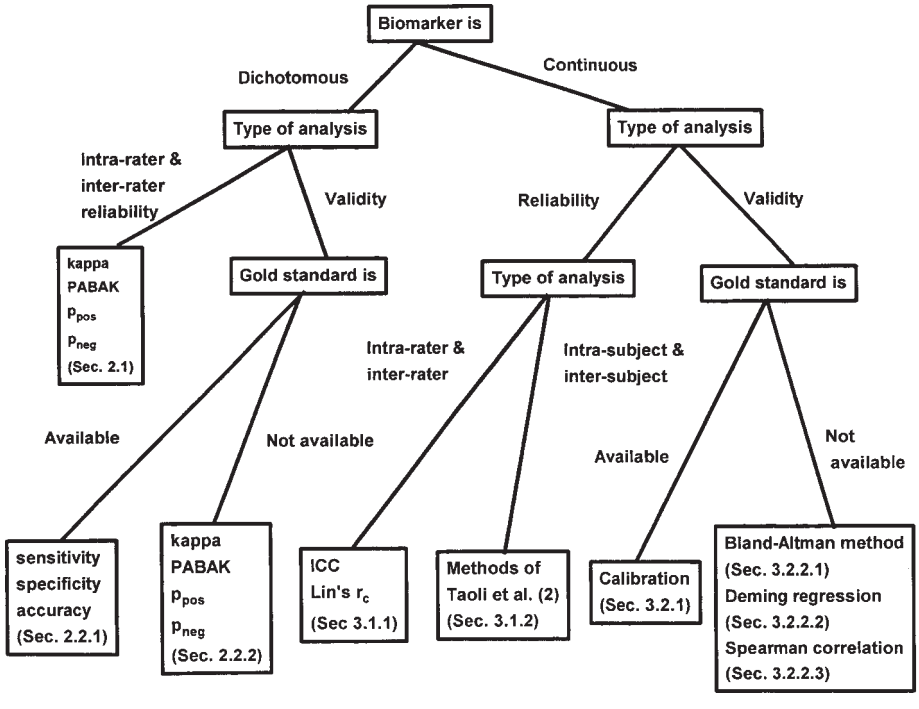

Fig. 4. Decision tree for determining the appropriate statistical method to use in assessing reliability or validity of a biomarker.

4. Discussion

In this chapter, we have described methods for assessing the reliability and validity of biomarkers that we feel are easy to apply and interpret, and whose results can be easily communicated to nonstatisticians. (**Figure 4** summarizes our recommendations in the form of a decision tree for easy reference.) Some of the methods we recommend are controversial; for example, there are those who claim that the adjustment for chance agreement in the calculation of Cohen's kappa is inappropriate when measuring the agreement between clinical observers ***(9)*** and that the Index of Crude Agreement is the correct measure to use. However, as the use of Cohen's kappa is so widespread and no one has come forward as yet with a convincing argument that kappa should be abandoned entirely, we have chosen to recommend its use, along with the PABAK coefficient and the indices of positive and negative agreement.

Our recommendation against the use of the intraclass correlation coefficient for assessing consistency between competing biomarkers may surprise some readers who are experienced with biomarker evaluation; however, we feel that

the criticisms by Bland and Altman ***(25)*** and Atkinson ***(26)*** are valid and that the methods recommended by Bland and Altman ***(16,21)***, and Strike ***(23)*** are preferable. Our recommendations are also consistent with the decision tree for the proper use of ICCs presented in ***(38)***.

We may also surprise some experienced readers by not recommending that significance tests be performed for κ, PABAK, the intraclass correlation ρ_I, and Spearman's rho. As Kraemer ***(39)*** and Altman and Bland ***(16)*** have pointed out, testing H_o: $\kappa = 0$ or H_o: $\rho_I = 0$ is beside the point because it is unlikely that we would be interested in the agreement between totally unrelated quantities in an assessment of reliability or validity. We prefer to use the guidelines provided by various authors as descriptors of the degree of agreement. Of course, these guidelines were not intended to be applicable in every situation and could be modified as necessary for the particular area of study. If a test of significance or confidence interval is required, the software used to calculate the coefficient can be used to produce these results as well (see below for software recommendations).

The coefficient of variation (CV) is commonly used as a measure of variability within assays, between assays, within samples, within individuals, etc. (*see*, e.g., ***[35,40]***). However, there are difficulties with the interpretation of the CV as it is usually presented (e.g., "the assay has a CV of 8%"). Strike (***23***, p. 25) provides a very lucid discussion of these difficulties and we agree with his recommendation: "Use the CV if you must, but use it carefully, and with proper qualifications."

In terms of software requirements for carrying out the procedures we have recommended, either the KAPPA or PAIRS programs of the software package PEPI ***(41)*** can be used to perform any of the calculations described here, with the exception of the intraclass correlation coefficient (ICC) in Eq. (3.1) and Deming Regression. PEPI is very reasonably priced (currently $50) and can be obtained from USD Inc., 2171-F West Park Ct., Stone Mountain, GA 30087, telephone (770) 469-4098, website www.usd-inc.com. The ICC in Eq. (3.1) can be calculated using the SAS code provided in ***(24)***, and the MINISNAP software provided with **ref.** ***(23)*** can be used to perform Deming regression.

Finally, an important issue that we have not addressed in this chapter is the assessment of surrogate markers that are used in place of a dichotomous event (death, recurrence of disease, etc.) as outcomes in clinical trials ***(42–45)***. Although many of the methods outlined in this chapter could be used to assess the reliability and validity of a surrogate marker, there are many other issues dealing with the use of these markers that are as yet unresolved, as illustrated by the recent "debate" at a workshop sponsored by the National Institute of Allergy and Infectious Diseases ***(45)***. A discussion of these issues is beyond the scope of this chapter. However, Chapter 9 of this text *Statistical Considerations in Assessing Molecular Markers for Cancer Prognosis and Treatment Efficacy* by Dignam et al. does consider some of the issues in detail.

Acknowledgment

I wish to thank Stephen George of the Duke University Medical Center for his many helpful comments that greatly improved this chapter. I also wish to thank Fred Benz of the University of Louisville School of Medicine for his encouragement and his informative presentations.

References

1. Last, J. M. (1995) *A Dictionary of Epidemiology*, (3rd edit.). Oxford University Press, New York.
2. Taioli, E., Kinney, P., Zhitkovich, A., Fulton, H., et al. (1994) Application of reliability models to studies of biomaker validation. *Environ. Health Perspect.* **102,** 306–309.
3. Qiao, Y-L., Tockman, M. S., Li, L., Erozan, Y. S., et al. (1997) A case-cohort study of an early biomarker of lung cancer in a screening cohort of Yunnan tin miners in China. *Cancer Epidemiol. Biomarker Prev*. **6,** 893–900.
4. Benowitz, L. (1999) Biomarkers of environmental tobacco smoke exposure. *Environ. Health Perspect.* **107**(Suppl 2)**,** 349–355.
5. Tockman, M. S., Gupta, P. K., Myers, J. D., Frost, J. K., et al. (1988) Sensitive and specific monoclonal antibody recognition of human lung cancer antigen on preserved sputum cells: a new approach to early lung cancer detection. *J. Clin. Oncol.* **6,** 1685–1693.
6. Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20,** 37–46.
7. Landis, J. R. and Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics* **33,** 159–174.
8. Bartko, J. J. (1991) Measurement and reliability: statistical thinking considerations. *Schizophr. Bull.* **17,** 483–489.
9. Feinstein, A. R. and Cicchetti, D. V. (1990) High agreement but low kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* **43,** 543–549.
10. Byrt, T., Bishop, J., and Carlin, J. B. (1993) Bias, prevalence, and kappa. *J. Clin. Epidemiol.* **46,** 423–429.
11. Cicchetti, D. V. and Feinstein, A. R. (1990) High agreement but low kappa: II. Resolving the paradoxes. *J. Clin. Epidemiol.* **43,** 551–558.
12. de Bock, G. H., Houwing-Duistermaat, J. J., Springer, M. P., Kievit, J., and van Houwelingen, J. C. (1994) Sensitivity and specificity of diagnostic tests in acute maxillary sinusitis determined by maximum likelihood in the absence of an external standard. *J. Clin. Epidemiol.* **47,** 1343–1352.
13. Joseph, L., Gyorkos, T. W., and Coupal, L. (1995) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am. J. Epidemiol.* **141,** 263–272.
14. Hui, S. L. and Zhou, X. H. (1998) Evaluation of diagnostic tests without gold standards. *Statist. Methods Med. Res.* **7,** 354–370.
15. Westgard, J. O. and Hunt, M. R. (1973) Use and interpretation of common statistical tests in method-comparison studies. *Clin. Chem.* **19,** 49–57.

16. Altman, D. G. and Bland, J. M. (1983) Measurement in medicine: the analysis of method comparison studies. *Statistician* **32,** 307–317.
17. Shrout, P. E. and Fleiss, J. L. (1979) Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* **86,** 420–428.
18. Fleiss, J. L. (1986) *The Design and Analysis of Clinical Experiments*. John Wiley & Sons, New York.
19. Bernstein, C., Bernstein, H., Garewal, H., Dinning, P., et al. (1999) A bile acid-induced apoptosis assay for colon cancer risk and associated quality control studies. *Cancer Res,* **59,** 2353–2357.
20. Lin, L. I. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45,** 255–268.
21. Bland, J. M. and Altman, D. G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* Feb. 8, 307–310.
22. Strike, P. W. (1991) *Statistical Methods in Laboratory Medicine*. Butterworth-Heinemann, Oxford.
23. Strike, P. W. (1996) Assay method comparison studies, in *Measurement in Laboratory Medicine: A Primer on Control and Interpretation*. Butterworth-Heinemann, Oxford, pp. 147–172.
24. Lee, J., Koh, D., and Ong, C. N. (1989) Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Comput. Biol. Med.* **19,** 61–70.
25. Bland, J. M. and Altman, D. G. (1990) A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput. Biol. Med.* **20,** 337–340.
26. Atkinson, G. (1995) A comparison of statistical methods for assessing measurement repeatability in ergonomics research, in *Sport, Leisure and Ergonomics* (Atkinson, G., and Reilly, T., eds.), E. and F. N. Spon., London, pp. 218–222.
27. Atkinson, G. and Nevill, A. (1997) Comment on the use of concordance correlation to assess the agreement between two variables (Letter to the Editor). *Biometrics* **53,** 775–777.
28. Lin, L. I. and Chinchilli, V. (1997) Rejoinder to the Letter to the Editor from Atkinson and Nevill. *Biometrics* **53,** 777–778.
29. Bartczak, A., Kline, S. A., Yu, R., Weisel, C. P., et al. (1994) Evaluation of assays for the identification and quantitation of muconic acid, a benzene metabolite in human urine. *J. Toxicol. Environ. Health* **42,** 245–258.
30. Kummel, C. H. (1879) Reduction of observation equations which contain more than one observed quantity. *Analyst* **6,** 97–105.
31. Linnet, K. (1990) Estimation of the linear relationship between the measurements of two methods with proportional errors. *Statist. Med.* **9,** 1463–1473.
32. Linnet, K. (1993) Evaluation of regression procedures for methods comparison studies. *Clin. Chem.* **39,** 424–432.
33. Coultas, D. B., Samet, J. M., McCarthy, J. F., and Spengler, J. D. (1990) A personal monitoring study to assess workplace exposure to environmental tobacco smoke. *Am. J. Public Health* **80,** 988–990.

34. Morton, R. F., Hebel, J. R., and McCarter, R. J. (1996) *A Study Guide to Epidemiology and Biostatistics*. Aspen, Gaithersburg, MD.
35. Amorim, L. C. A. and Alvarez-Leite, E. M. (1997) Determination of *o*-cresol by gas chromatography and comparison with hippuric acid levels in urine samples of individuals exposed to toluene. *J. Toxicol. Environ. Health* **50,** 401–407
36. Matsunga, S. K., Plezia, P. M., Karol, M. D., Katz, M. D., et al. (1989) Effects of passive smoking on theophylline clearance. *Clin. Pharmacol. Ther.* **46,** 399–407.
37. Hüttner, E., Götze, A., and Nikolova, T. (1999) Chromosomal aberrations in humans as genetic endpoints to assess the impact of pollution. *Mutat. Res.* **445,** 251–257.
38. Müller, R. and Büttner, P. (1994) A critical discussion of intraclass correlation coefficients. *Statist. Med.* **13,** 2465–2476.
39. Kraemer, H. C. (1980) Extension of the kappa coefficient. *Biometrics* **36,** 207–216.
40. Atawodi, S. E., Lea, S., Nyberg, F., Mukeria, A., et al. (1998) 4-Hydroxyl-1-(3-pyridyl)-1-butanone-hemoglobin adducts as biomarkers of exposure to tobacco smoke: validation of a method to be used in multicenter studies. *Cancer Epidemiol. Biomark. Prev.* **7,** 817–821.
41. Abramson, J. H. and Gahlinger, P. M. (1999) *Computer Programs for Epidemiologists, PEPI Version 3.00*. Brixton Books, Llanidloes, Powys.
42. Fleming, T. R., Prentice, R. L., Pepe, M. S., and Glidden, D. (1994) Surrogate and auxillary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statist. Med.* **13,** 955–968.
43. Fleming, T. R. (1994) Surrogate markers in AIDS and cancer trials. *Statist. Med.* **13,** 1423–1435.
44. Buyse, M. and Molenberghs, G. (1998) Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* **54,** 1014–1029.
45. Albert, J. M., Ioannidis, J. P. A., Reichelderfer, P., Conway, B., et al. (1998) Statistical issues for HIV surrogate endpoints: Point/counterpoint. *Statist. Med.* **17,** 2435–2462.