Using the Coefficient of Correlation in Method-Comparison Studies

Michael J. Bookbinder and Kim J. Panosian

The coefficient of correlation (R) is one of the most commonly computed statistics in method-comparison studies. Usually, it is simply quoted without interpretation. In this paper, we show how R may be used to detect interference, nonlinearity, and misuse of the imprecision components. Specifically, one may precisely predict what R should be by considering the imprecisions of the two methods being compared, even before the comparison is performed. When the actual R disagrees with the predicted R, then one of the mentioned effects is present. We also describe a statistical test to detect these effects at the P = 0.05 level, then evaluate this test by using computer simulation and present two examples of its use. We also present the theory underlying the usage of R, including how R is affected by the distribution and range of the data, by the joint imprecisions of the methods being compared, by the sample size, and by the randomness of the specimen-selection process.

Additional Keyphrases: statistics · variation, source of · interferences

Perhaps the most common method of assessing accuracy is the method-comparison study. In such studies (1, 2), at least 40 (3) different specimens are measured by both a comparison method and the method in question. When the results of the two methods are plotted against each other, perfect agreement would yield the straight line y = x. Thus, bias is measured by the distance from this line to the least-squares line actually obtained.

Investigators making a method-comparison study usually compute the coefficient of correlation (R), even though it is insensitive to inaccuracy and is affected by the range of the specimens (4). In fact, the expected value of R may be computed (see below) from the precision of the two methods if they are linearly related. Graphically, the scatter of points about the least-squares line reflects the joint imprecision, and the location of this line reflects relative bias (Figure 1). However, imprecision may be assessed in the usual manner by measuring 20 to 30 aliquots of the same specimen and computing the coefficient of variation (CV). Because R does measure imprecision (when corrected for the range), there should be agreement between R and the precision of the two methods: R should be close to 1.0 when both methods are precise.

However, we will show that R may be low even when the precisions of both methods are good. When the imprecision as estimated by aliquot measurement disagrees with the imprecision implied by R, there must be some breakdown in the assumptions underlying the method-comparison study. If a way of predicting R from the two precisions can be devised, then the difference between the predicted R and the actual R would indicate the amount of extraneous influence present.

Here we present a method for predicting R, given the pre-existing precisions of the two methods, and we describe a statistical test that may be used to test whether the actual R differs from this expected R at the P = 0.05 level. Such a test may be used to demonstrate interference, which may not be detected by either precision analysis or bias estimation. The test also allows detection of other problems such as nonlinearity. Two examples are presented, including a situation in which the actual R is low even though the joint precision of the two methods is good.

Method for Calculating the Population R

From statistical theory, the value of the population R can be estimated if the imprecision of the two methods is known and if certain assumptions are valid (see Appendix 1). In particular, if the imprecision of the two methods can be represented as a constant imprecision over the relevant range (e.g., 2 mg/L for method 1 and 4 mg/L for method 2), then the estimated value of R will be given by:

$$ R = \sqrt{1 - \frac{\sigma_1^2}{\text{var}(x)}} \left[1 - \frac{\sigma_2^2}{\text{var}(y)}\right] $$

(1)

$\sigma_1^2$ is the constant imprecision of method 1 expressed as a variance, and var(x) is the variance of the specimens as measured by method 1. $\sigma_2^2$ is the imprecision of method 2 and var(y) is the variance of the specimens as measured by method 2. For example, $\sigma_1^2$ could be the variance of 20 measurements of a specimen having a representative mean value.

When the imprecision of the two methods is best represented by some proportionate error (a constant coefficient of variation), then:

$$ R = \sqrt{\frac{1 - [(CV_1^2)(\sigma_2^2)/\text{var}(x)]}{(1 + CV_1^2)}} \left[1 - [(CV_2^2)(\mu_2^2)/\text{var}(y)]\right] $$

(2)

Here, CV_1 and CV_2 are the constant coefficients of variation of methods 1 and 2, respectively, expressed as a decimal (e.g., 0.02 = 2%) and $\sigma_1$ and $\sigma_2$ are the mean and variance of the specimens as measured by method 1, and similarly for $\mu_2$, var(y), and method 2.

The imprecisions used in equations 1 and 2 must be within-day imprecisions if the specimens are all measured within one day. Between-day imprecisions should be used if the specimens are measured once a day for several days.

While a CV tends to be dependent upon the analyte concentration, often a constant standard deviation will approximate the analytical error at most concentrations. Thus equation 1 will be used more often than equation 2.

These equations express mathematically what Westgard...
and Hunt found empirically (4): R is a function of not only the imprecision but also the range of the observed results [var(x) or var(y)]. These equations are independent of the distribution of the specimens and of the distribution of the error in reading the data.

These equations allow estimation of the population R for readings from two assays having a linear bias between them if their imprecisions are known. The sample R, however, is the only R that is obtainable, and it will depend upon (a) the number of specimens, (b) the distribution of the specimens, and (c) the value of the population R (from equation 1 or 2). However, a method for testing the sample R for consistency with the imprecision of the two methods can be constructed.

Testing R Derived from Randomly Sampled, Normally Distributed Data

The actual R obtained may be compared with the R predicted by equation 1 or 2 by constructing a 95% confidence interval for R, based on certain assumptions. If the sample R lies outside the 95% interval, one of the assumptions has been violated. The assumptions are as follows: (a) the two methods are linearly related—the readings of one method are in theory equal to those of the other method after a multiplicative and an additive correction; (b) the data are randomly selected from an underlying normal population; and (c) no interference (see Appendix 2) is present.

The data from one of the two methods may be shown to be normally distributed either visually or by the Kolmogorov–Smirnov test (5–8), but some deviation from a normal curve is permissible (see Appendix 3). It is also not necessary to show that the data from both methods are normal; either one is sufficient (see Appendix 3). One then computes the confidence interval for R by using the Fisher method:

- Obtain the expected population R from equation 1 or 2.
- Convert this R to the Fisher variate, z, by:
  \[ z = 0.5 \cdot \log_a((1 + R)/(1 - R)) \]
- Compute the standard deviation of z, \( \sigma_z \), by using \( \sigma_z = 1/\sqrt{(N - 3)} \), where N is the number of specimens in the study.
- Compute the upper and lower limits on z as:
  \[ \text{Upper limit} = z + 1.96 \cdot \sigma_z \]
  \[ \text{Lower limit} = z - 1.96 \cdot \sigma_z \]
- Convert these two limits into two R values by using the inverse formulas:
  \[ R_{UL} = [\exp(2 \cdot UL) - 1]/[\exp(2 \cdot UL) + 1] \]
  \[ R_{LL} = [\exp(2 \cdot LL) - 1]/[\exp(2 \cdot LL) + 1] \]
  where \( \exp(x) = e^x \).

The two Rs so obtained (\( R_{UL} \) and \( R_{LL} \)) define the 95% confidence limits on the sample R obtained from the comparison study. A correction factor may be needed (see below).

The above method is known to be valid for calculating the 95% confidence interval for the sample R when drawing from a bivariate normal population. However, because method-comparison studies may not involve bivariate normal populations (see Appendix 3), the validity of the method must be established. We used computer simulation to evaluate the usefulness of the above methodology in predicting R in a method-comparison study. As shown in Table 1, if the data from one of the methods is normally distributed, then the above method gives the expected value of R accurately. The 95% range of R is quite accurate at high values of R, and this range is fairly accurate at low values of R.

One may make the 95% range of R more accurate by multiplying the predicted lower limit, \( R_{LL} \), by 1 + 0.05 \cdot (1 - \( R_{LL} \)), an adjustment we derived empirically using computer simulation. For example, in Table 1, if the predicted R were 0.831, one would calculate a rough interval for R as 0.701–0.908. This could be improved by multiplying the lower limit, 0.701, by 1 + 0.05 \cdot (1 - 0.701). The value 0.711 so obtained is closer to the true lower limit of 0.713. This correction is not needed if the data are distributed as

<table>
<thead>
<tr>
<th>CV1, %</th>
<th>CV2, %</th>
<th>Obtained by simulation</th>
<th>Predicted by eq. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.997</td>
<td>.997*</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>.989</td>
<td>.989*</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>.976</td>
<td>.976*</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>.958</td>
<td>.958*</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>.934</td>
<td>.934*</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>.905</td>
<td>.905*</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>.833</td>
<td>.833*</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>.741</td>
<td>.741*</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>.919</td>
<td>.919*</td>
</tr>
</tbody>
</table>

\*Number of specimens = 40. The mean and SD of the specimens were approximately 200 and 38.5, respectively, for both methods. In the simulations of the last line, the specimen SD (method 1) was 42.5.* See Appendix 3. 

This interval can be improved by multiplying the lower limit, \( R_{LL} \), by 1 + 0.05 \cdot (1 - \( R_{LL} \)). The ranges shown have not been so corrected.

Table 1. Agreement of Predicted R and 95% Confidence Interval with Results of Computer Simulation* for Normally Distributed Specimens

CLINICAL CHEMISTRY, Vol. 33, No. 7, 1987 1171
Simulation studies showed that the model works equally well with a constant analytical error and that the 0.05 \( \cdot (1 - R_{LL}) \) correction may be used for that case also.

Testing \( R \) Derived from Data Randomly Selected from a Bimodal Population

Often it is desirable to compare two methods by using data from a population consisting of mostly normal subjects admixed with some abnormal subjects. In this case, the data may not be normally distributed, but might have a bimodal distribution. We therefore evaluated whether the above method would work when patients are randomly selected from a bimodal population. We used a Gram–Charlier Curve with \( \gamma \) equal to 2 (Figure 2) to generate the distribution of the "true" specimen values. We then used computer simulation to "measure" these values with analytical error. Equations 1 and 2 predicted \( R \) well and the calculated interval for \( R \) was accurate (Table 2). Moreover, the 1 + 0.05 \( \cdot (1 - R_{LL}) \) correction was not needed. Apparently, \( R \) is robust enough to be predicted from populations that deviate significantly from normal (at \( N \approx 40 \)).

Testing \( R \) Derived from Nonrandom Data

For many analytes, selecting specimens at random will lead to an approximately normal distribution. However, it is sometimes desirable to examine accuracy over a broad range. In this case, specimens are not selected randomly but are chosen to include a wide range of values (3). We used computer simulation to examine a case in which the underlying distribution of specimens was normal, but when enough specimens were collected near the mean, random specimens were rejected until more specimens were identified as outlying. We required that at least 15% of the specimens measured by method 1 be 2 SD or more from the mean (usually, only 4.6% of specimens should be out this far).

The results of the simulations are shown in Table 3. The absolute values of \( R \) (from the simulations, Table 3) were higher than those obtained from purely random sampling (Table 1). This increase probably occurred because the range of specimens (var(x) and var(y)) was forced to be wide. However, equation 2 still predicted \( R \) almost exactly.

The predicted range of \( R \) was slightly wider than the range given by the simulator, even with the 5% adjustment. Selecting specimens at the extremes seemed to force \( R \) to be higher and less variable. Because this error is conservative and small after adjustment, the adjusted Fisher interval may be used as a guide for testing \( R \) in this case also.

Two Examples

We first compared two methods not expected to have interferences: Glucose readings for 44 consecutive patients were obtained via the ASTRA 8 (Beckman Instruments, Brea, CA) and the S-MAC I (Technicon Corp., Tarrytown, NY). The specimens were all analyzed within one run with each

---

### Table 2. Agreement of Predicted \( R \) and 95% Confidence Interval with Results of Computer Simulation* for Gram–Charlier-Distributed Specimens

<table>
<thead>
<tr>
<th>CV1, %</th>
<th>CV2, %</th>
<th>( R )</th>
<th>95% Interval</th>
<th>Predicted by eq. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.998</td>
<td>.996–999</td>
<td>.998</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>.991</td>
<td>.984–996</td>
<td>.991</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>.981</td>
<td>.964–990</td>
<td>.981</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>.965</td>
<td>.936–983</td>
<td>.965</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>.946</td>
<td>.899–973</td>
<td>.946</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>.921</td>
<td>.856–952</td>
<td>.921</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>.859</td>
<td>.746–929</td>
<td>.858</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>.775</td>
<td>.611–886</td>
<td>.775</td>
</tr>
</tbody>
</table>

*Number of specimens = 40. Mean and SD of the specimens were approximately 196.3 and 43, respectively. In the simulations of the last line, the SD of the specimens was 46 for method 1. With random analytical error added.

### Table 3. \( R \) and Its Range as Predicted by Simulation When Samples Are Selected to Lie over a Broadened Range*

<table>
<thead>
<tr>
<th>CV1, %</th>
<th>CV2, %</th>
<th>( R )</th>
<th>95% Interval</th>
<th>Predicted by eq. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.998</td>
<td>.997–999</td>
<td>.998</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>.993</td>
<td>.988–996</td>
<td>.993</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>.984</td>
<td>.974–991</td>
<td>.984</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>.972</td>
<td>.953–985</td>
<td>.971</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>.955</td>
<td>.926–976</td>
<td>.954</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>.934</td>
<td>.892–964</td>
<td>.933</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>.880</td>
<td>.804–934</td>
<td>.879</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>.803</td>
<td>.777–894</td>
<td>.805</td>
</tr>
</tbody>
</table>

* Fifteen percent of specimens were required to be 2 SD or more away from the mean, but underlying population distribution is normal. Number of specimens = 40. Mean and SD of specimens finally selected were approximately 200 and 48, respectively, for both methods; however, the specimen SD for method 2 tended to be somewhat lower at the higher CVs and in the simulations of the last line. Standard deviation of underlying population was 40 for both methods. * This interval can be improved by multiplying the lower limit, \( R_{LL} \), by 1 + 0.05(1 – \( R_{LL} \)).

---

1172 CLINICAL CHEMISTRY, Vol. 33, No. 7, 1987
Specimens all measured within one run so that scatter reflects joint within-run imprecision.

The actual R was 0.998 (Figure 3). The SMAC was known to have a precision of roughly \( \sigma_{WR} = 46 \text{ mg/L} \) over the entire range. The Astra's precision was approximately 20 mg/L. The 44 specimens had the following dispersion:

- **Astra**:
  - \( \text{mean} = 1580, \ SD = 800 \ (N = 44) \)
  - \( \text{SD} = 840 \ (N = 44) \)

Using equation 1, the expected R is:

\[
R = \sqrt{1 - \frac{(20)^2}{(800)^2} \cdot \frac{1 - (46)^2}{(840)^2}} = 0.998
\]

Note the excellent agreement between the predicted R and the actual R. Next, the 95% confidence limits on the sample R were computed to be 0.997–0.999. Since the obtained R, 0.998, falls within this interval, no interferences were detected by this test. No correction factor is needed for large R values.

We then compared a method known to be subject to an interference with one that is not. Forty specimens, about half known to have a high lipid content, were analyzed for phosphate with the SMAC I and the AHA (DuPont Instruments, Wilmington, DE). The specimens were selected randomly with respect to phosphate concentration. All 40 specimens were analyzed within one run so that only within-run imprecision would be involved. The within-run precision of both instruments was best represented by a constant standard deviation rather than a standard deviation proportional to the concentration of phosphate. The results are shown in Figure 4. The within-run standard deviation of the AHA was 0.35 mg/L at 34 mg/L and 0.44 mg/L at 59 mg/L. We used a constant standard deviation of 0.4 mg/L. The SMAC's precision was such that the standard deviation was 1.4, 1.0, 0.9, and 1.7 mg/L at 12, 30, 34, and 57 mg/L, respectively. For the SMAC I we used 1.2 as the approximately constant standard deviation over the entire range of phosphate concentrations.

Analysis of the 40 data points yielded an R of 0.879. The variance of the 40 measurements was \( \sigma^2(\text{AHA}) = 347.27 \) and \( \sigma^2(\text{SMAC}) = 189.48 \text{ (mg/L)}^2 \). By equation 1:

\[
R = \sqrt{1 - \frac{(1.2)^2}{189.48} \cdot \frac{(0.4)^2}{347.27}} = 0.996
\]

R = 0.996 is the population R expected given the two levels of precisions. Using the Kolmogorov–Smirnov test with the AHA data indicated that a normal distribution was consistent with the data (7, 8) (\( D = 0.132 \)). To compute the 95% confidence interval for the sample R, we obtained:

- **Upper limit** = \( z + 1.96 \cdot \sigma_z = 3.4285 \)
- **Lower limit** = \( z - 1.96 \cdot \sigma_z = 2.7841 \)

By the inverse formula, \( R_{UL} = 0.998 \) and \( R_{LL} = 0.992 \). Thus, in this example involving a marked lipid interference, the sample R obtained, 0.879, lies very much outside the 95% confidence interval (0.992–0.998).

**Interpretation**

In a method-comparison study, a high degree of correlation is already expected so that the absolute value of R (or of the Spearman rank correlation coefficient (9)) is of little use by itself. However, the method of this paper allows the actual R to be compared with the R predicted from equation 1 or 2. When the actual R is outside the confidence interval, one of the following violations of the assumptions may be present:

1. An interfering substance is present. However, our method will not indicate which method is subject to the
interference unless one of the methods is known to be interference-free. Note that the converse is not true: obtaining an $R$ within the 95% confidence interval does not imply that a less pervasive but still significant interference is not present. Interferences must still be systematically sought after. However, the tendency of $R$ to be strongly influenced by extreme values (4) indicates that $R$ may be sensitive to infrequent interferences. The range also affects $R$, but this does not affect the validity of our method, because the predicted $R$ is already adjusted for the range (equations 1 and 2).

2. The two imprecisions are not being used properly. As mentioned above, the predicted $R$ (equations 1 and 2) must be calculated from imprecisions that reflect how the specimens are measured. For example, if samples are measured over several days, then $R$ will reflect the total (day-to-day) CV as opposed to the within-day CV. However, in equations 1 and 2, day-to-day variation may be used if only one specimen is measured per day. When more than one (but less than all) specimens are measured per day, some intermediate CV will be applicable (10). Generally, measuring all the samples within one run will give a tighter fit about the least-squares line with a better $R$; however, analytical error that varies over time will not be detected.

3. There is a nonlinear relationship between the two tests. That is, bias cannot be corrected by a multiplier plus an added constant.

4. The distribution of the data is unusual. If the shape of the plotted data is very different from both normal and Gram–Charlier curves, predicting $R$ as described above may be inaccurate.

Effect of Interference on Bias and Imprecision

In a method-comparison study, we hypothesize that interference may produce effects similar to either bias or inaccuracy, as follows. If the interference acts in one direction only (e.g., sometimes making results by one method too high), then the interference must be sporadic or variable in its effect for it to be recognized as an interference. Otherwise, a uniform bias would simply appear to be present.

When an interference acts (sporadically) in one direction, the methods will still appear to be relatively biased by an additional amount due to the interference. However, scatter about the least-squares line will also be present beyond that expected from the two precisions (see Figure 1D). If mixed interferences are present, sometimes increasing the reading for one of the assays and other times decreasing it, then the two methods will show an unexpected scatter about the least-squares line (Figure 1B). There may or may not appear to be additional bias due to this interference. The obtained $R$ should be lower than what is predicted by equation 1 or 2.

In summary: the coefficient of correlation can be readily predicted from the imprecision of the two methods being compared. We have also shown how a 95% confidence interval can be determined for the sample $R$. Although the absolute value of $R$ is not useful in itself, comparing it with the expected $R$ by using this confidence interval makes possible meaningful interpretation. This test is useful for detecting significant interferences by an aggregate analysis. It explains why method-comparison graphs may have greater scatter than expected from the joint precisions of the two methods. In addition to scatter from interfering substances, measuring samples over several days will lead to a scatter (or $R$) that is not consistent with the within-day imprecisions of the two methods. If the relative bias between the two tests is nonlinear, then the method may not work. The limitation of the method is that obtaining the expected $R$ still does not exclude a less-pervasive interference. It will not detect an interference that affects both methods identically. The method may be inaccurate if the patients’ data are radically distributed or are selected very nonrandomly. On the other hand, understanding the principles of this method should be useful for guiding laboratory workers in the use and interpretation of this extensively used statistic.

We are grateful to Martha Carey, Lesley Orlovski, and Paul McNamara for their technical assistance.

References


Appendix I. Statistical Theory for Predicting the Population $R$ from the Imprecision

The following derivation comes from the theory of linear functional relationships and of measurement error (11, 12). This model is also the correct one to use for properly calculating the least-squares line for estimating bias in method-comparison studies (13–15).

Case 1: Constant imprecision (e.g., $\sigma^2 = 5$ mg/L for method 1). If two methods, 1 and 2, are being compared,
each specimen will have the theoretical value X and Y for methods 1 and 2, respectively. If there is no more than a constant plus a proportionate bias between the two methods, then \( Y = aX + b \). Furthermore, assume that the true values of any specimen, X and Y, can never be observed; rather, X and Y can only be measured, with error, by the two respective assays. If the observed readings of the specimens are \( x \) and \( y \), then \( x = X + e_1 \) and \( y = Y + e_2 \) where \( e_1 \) and \( e_2 \) are normally occurring (analytical) errors with an expected value of zero. The variance of \( e_1, \sigma^2_1 \), is the constant imprecision of method 1, and the variance of \( e_2, \sigma^2_2 \), is the constant imprecision of method 2. No assumptions need be made about the distribution of the \( e_i \); rather, we assume that there are no sporadic (interfering) effects on measurements.

It follows from the above that \( \text{var}(x) = \text{var}(X) + \sigma^2_1 \) and \( \text{var}(y) = \text{var}(Y) + \sigma^2_2 \). Also \( \text{cov}(x,y) = \text{cov}(X,Y) = a \cdot \text{var}(X) \).

By definition, the coefficient of correlation between the observed variables, \( x \) and \( y \), is given by:

\[
R(x,y) = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}}
\]

Substituting into this we obtain:

\[
R(x,y) = \frac{a \cdot \text{var}(X)}{\sqrt{\text{var}(X) + \sigma^2_1[\text{var}(Y) + \sigma^2_2]}}
\]

Substituting \( a^2 \cdot \text{var}(X) = \text{var}(Y) \), then:

\[
R = \frac{1}{\sqrt{1 + \frac{\sigma^2_1}{\text{var}(X)}} \left[ 1 + \frac{\sigma^2_2}{a^2 \cdot \text{var}(X)} \right]}
\]

The first equation may be solved to yield:

\[
\text{var}(X) = \frac{\text{var}(x) - (CV_1 \cdot \mu_2)^2}{(1 + CV_1^2)}
\]

A similar equation may be found for \( \text{var}(Y) \). Substituting the first three equations into the definition of \( R \):

\[
R(x,y) = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}}
\]

we obtain:

\[
R(x,y) = \frac{a \cdot \text{var}(X)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}
\]

where product =

\[
\{\text{var}(X) + \sigma^2_1[\text{var}(X) + \mu_2^2]\} \cdot \{\text{var}(Y) + \sigma^2_2[\text{var}(Y) + \mu_2^2]\}
\]

or, product = \( \{\text{var}(X) [1 + \sigma^2_1]\}

\[
+ \text{var}(X) \cdot \mu_2^2[a^2 \cdot \text{var}(X)[1 + \sigma^2_2] + \text{var}(Y) \cdot \mu_2^2]
\]

Dividing by \( a \cdot \text{var}(X) \), we obtain

\[
R = \frac{1}{\sqrt{1 + \frac{\sigma^2_1}{\text{var}(X)} + \frac{(\mu_2 \cdot \sigma_2)^2}{\text{var}(X)}[1 + \frac{\sigma^2_2}{(a^2 \cdot \text{var}(X))^2]}}}
\]

If we substitute the expressions for \( \text{var}(X) \) and \( (Y) \) into this equation, we obtain equation 2 after rearranging. Here \( \sigma_1 \) and \( \sigma_2 \) are equal to CV1 and CV2, respectively.

**Appendix 2: Interferences**

We used an operational definition of interference to refer to any effect that alters the results of one method relative to another beyond what is expected from the linear bias and the imprecision of the methods. Interference might then be detected by the method presented in this paper by assaying multiple samples, or by using a one-point method. That is, interference exists in one method relative to another if two specimens can be found, for example, that yield the following results upon repeated measurement:

**Appendix 3: Statistical Notes on Distributions**

If the data are normally distributed as measured by one of the assays, the data from the other method should be normally distributed also, if the methods are linearly related without interference. One of these assumptions must be incorrect if the data distribution is normal by one method but not by the other. When this happens, we recommend proceeding with the calculations to verify that this effect is significant (according to the 95% confidence interval).

However, even when the data from both methods are normal in shape, a bivariate normal situation will not occur, for example, if the analytical error is proportionate. In a bivariate normal distribution, the conditional variance or
error of $Y$ about $X$ must be constant.

The fact that the underlying population is not bivariate normal does not invalidate the use of $R$. The properties and applications of $R$ when drawn from nonnormal populations have been extensively studied (16).

Finally, given an underlying but unobservable specimen population that is normally distributed, the observed specimens, when having a constant normally distributed error, will be normally distributed because the sum of two independent normals is normal. However, if the error is proportionate, the observable data will be only approximately normal, owing to the multiplicative effect of the analytical error. Probably, observable data will be distributed with this distortion if there is proportionate error.

All of the simulations involving proportionate error (Tables 1, 2, and 3) incorporated this distortion. As we mentioned in the text, the model predicted $R$ and its range well. Note that simulations involving a constant analytical error (and hence no distortion of the normal curve) produced the same valid results.