Statistical Inference for the Area under the Receiver Operating Characteristic Curve in the Presence of Random Measurement Error

Enrique F. Schisterman,1 David Faraggi,2 Benjamin Reiser,2 and Maurizio Trevisan1

The area under the receiver operating characteristic curve is the most commonly used measure of the ability of a biomarker to distinguish between two populations. Some markers are subject to substantial measurement error. Under normality assumptions, the authors develop a confidence interval procedure for the area under the receiver operating characteristic curve that adjusts for measurement error. This procedure assumes the availability of data from a reliability study of the biomarker. A simulation study was used to check the validity of the proposed confidence interval. Furthermore, it was shown that not adjusting for measurement error could result in a serious understatement of the effectiveness of the biomarker. Am J Epidemiol 2001;154:174–9.

cardiovascular diseases; reliability; repeated measures; thiobarbituric acid reaction substances

Receiver operating characteristic (ROC) analysis was originally developed for use with radar technology to separate observer variability from the innate detectability of the signal (1). It has gained popularity in laboratory medicine over the past 30 years as a tool for evaluating test performance. ROC curves simultaneously show the proportion of both “abnormal” and “normal” subjects correctly diagnosed at various test cutoff points. This graphic display not only facilitates the selection of an optimal threshold but also enables the easy comparison of different tests. The use of this tool became popular in epidemiologic research as a statistical instrument to evaluate the discriminative abilities of different biomarkers, such as oxidative stress and antioxidants biomarkers.

The area under the ROC curve, introduced to the medical literature by Hanley and McNeil (2), is a frequently used criterion to examine the discriminative effectiveness of biomarkers. With X and Y representing the biomarker values on the controls and cases, respectively, Bamber (3) showed that the area under the ROC curve is \( A = P(Y > X) \); thus, A is a global measure of how well the biomarker distinguishes between cases and controls. Assuming that X and Y follow the normal distribution, Lloyd (4) provided a point estimator for A, while Reiser and Guttman (5) developed confidence intervals.

Measurement error and biases may be attributed to laboratory equipment, variation between technicians, temporal changes, biologic variability, etc. Random measurement error can bias possible relations toward the null hypothesis (6). Many methods are available for correcting correlation and regression coefficients for random measurement error by using the reliability approach (7, 8). In the context of ROC curves, Coffin and Sukhatme (9, 10) provided a bias correction approximation for the point estimator of A under random measurement error in both parametric and nonparametric cases. Faraggi (11) considered the effect of random measurement error on the confidence interval for A in the context of the parametric normal model. He showed numerically that not taking measurement error into account can give seriously misleading results that underestimate the diagnostic effectiveness of the marker and result in the stated confidence intervals for A having actual coverage substantially less than their nominal values. Assuming normality, Reiser (12) developed a corrected confidence interval for A, adjusted for measurement error. This method assumes the availability of replicated observations on the study subjects. However, such replications are not frequently collected in epidemiologic research because of cost and feasibility considerations. In epidemiologic and medical research, reliability and validation studies are frequently used to assess the magnitude of different sources of error (7) on the exposure of interest.

In this paper, we present both an estimator and a confidence interval of the area under the ROC curve under normality assumptions in the presence of random normal measurement error when replications are not available in the main study and an external reliability study is performed to evaluate measurement error. In the section Statistical Inference in the Presence of Random Measurement Error, by using the Delta method (13), we develop an approximate confidence interval of the area when random measurement error is present. A simulation study that examines how the bias of the estimator and the coverage properties of the approximate confidence interval are influenced by different sample sizes and variances is discussed.
in the Simulation Study section. We provide details of the oxygen free radicals biomarker study and show the importance of our methodology, namely, we show how incorporating corrections for random measurement error affected the results.

STATISTICAL INFERENCE IN THE PRESENCE OF RANDOM MEASUREMENT ERROR

Assume that a biomarker follows the normal distribution $X_i \sim N(\mu_x, \sigma_x^2)$ ($i = 1, \ldots, n_x$) in the control population and the normal distribution $Y_i \sim N(\mu_y, \sigma_y^2)$ ($i = 1, \ldots, n_y$) in the case population. The area under the ROC curve in this case is $A = \Phi(\delta)$, where $\delta = \frac{\mu_y - \mu_x}{\sqrt{\sigma_x^2 + \sigma_y^2}}$ and $\Phi$ denotes the standard normal cumulative distribution function. The almost maximum likelihood estimate (5) of $\delta$ is 

$$\hat{\delta} = \Phi\left(\frac{\bar{Y} - \bar{X}}{\sqrt{S_x^2 + S_y^2}}\right),$$

where $\bar{X}, \bar{Y}, S_x^2,$ and $S_y^2$ denote the sample means and variances, for the two populations (controls and cases), respectively.

In the presence of measurement error, the “true” value of the biomarker on the case and control groups ($Y$ and $X$) is not available. Instead, we assume that we actually observe

$$x_i = X_i + \epsilon_i \quad i = 1, \ldots, n_x$$

$$y_i = Y_i + \epsilon'_i \quad i = 1, \ldots, n_y$$

where $\epsilon_i \sim N(0, \sigma_x^2)$ and $\epsilon'_i \sim N(0, \sigma_y^2)$. We further assume that the $X, Y, \epsilon_i,$ and $\epsilon'_i$ are independent of each other. In many situations, it is reasonable to assume that the variations leading to measurement error are due to factors connected with the laboratory or measurement process and do not depend on the person’s risk status or the true biomarker value. Consequently, the variances of $\epsilon_i$ and $\epsilon'_i$ are taken to be equal.

In addition, we assume the availability of data from a reliability study designed to estimate measurement error. If we let $w_{ij} i = 1, \ldots, n_x j = 1, \ldots, p_i$ denote the $j$th observation on the $i$th subject, we assume that

$$w_{ij} = W_{ij} + \epsilon_{ij},$$

where $W_{ij}$ is the “true” value of the biomarker for the $i$th subject of the reliability study and $\epsilon_{ij} \sim N(0, \sigma^2)$. Let $n_f = \sum_{i=1}^{n_x} (p_i - 1)$ and $\bar{W}_r = \frac{1}{n_f} \sum_{i=1}^{n_x} \sum_{j=1}^{p_i} w_{ij}$. Consequently,

$$\hat{\delta}_r = \frac{1}{n_f} \sum_{i=1}^{n_x} \sum_{j=1}^{p_i} (w_{ij} - \bar{W}_r)^2$$

is an unbiased estimator of $\sigma^2$.

Let $\bar{x}, \bar{y}$ denote the sample means for the observed $x_o$, $y_i$ and set $S_x^2 = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (x_i - \bar{x})^2$ and $S_y^2 = \frac{1}{n_y - 1} \sum_{i=1}^{n_y} (y_i - \bar{y})^2$.

Consequently from equation 1, since $E(S_x^2) = \sigma_x^2 + \sigma^2$, an unbiased estimator of $\sigma^2$ is $\tilde{\sigma}_x^2 = S_x^2 - \hat{\delta}^2_r$. Similarly, $\tilde{\sigma}_y^2 = S_y^2 - \hat{\delta}^2_r$ is unbiased for $\sigma^2_y$.

Thus, when random measurement error is present, a natural corrected for measurement error, estimate of $A$ is $\hat{\delta}_c = \Phi(\hat{\delta})$, where $\hat{\delta} = \frac{\bar{Y} - \bar{X}}{\sqrt{S_x^2 + S_y^2 - 2\hat{\delta}_r^2}}$.

Note that the $P(S_x^2 < \tilde{\sigma}_x^2)$ > 0 and $P(S_y^2 < \tilde{\sigma}_y^2)$ > 0, implying that $S_x^2 - \tilde{\sigma}_x^2$ or $S_y^2 - \tilde{\sigma}_y^2$ may be negative, which will yield $\hat{\sigma}_x^2 < 0$ or $\hat{\sigma}_y^2 < 0$ and an undefined $\hat{\delta}$. This situation is similar to the classical variance components problem discussed by Searle (14). We follow Puduri and Rao (15), who recommend replacing negative variance estimates by a very small number.

Consequently, taking measurement error into account results in a larger estimate of $\hat{\delta}$ and, correspondingly, of $A$. Ignoring measurement error will underestimate $\hat{\delta}$ and will therefore underestimate the effectiveness of biomarkers in discriminating between cases and controls.

The amount of measurement error needed to be determined relative to the inherent marker variability. A commonly used index, termed the “reliability index” in the social science literature (16), is extended here to the two-populations case. This natural extension results in $R = \frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2 + \sigma_y^2 + 2\hat{\delta}_r^2}$. This index ranges between 0 and 1, with values close to 1 indicating that the measurement error is relatively small and values close to 0 indicating a large relative measurement error. Related reliability indexes are used in somewhat different situations by Faraggi (11) and Reiser (12).

Since $\Phi$ is a monotonically increasing function of $\delta$, finding a confidence interval for $A$ is equivalent to finding one for $\delta$. Using the Delta method, we estimate the variance of $\hat{\delta}$ by

$$\text{V\text{\textbar}r(}\hat{\delta}) = \left[\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}\right] \times \left(S_x^2 + S_y^2 - 2\hat{\delta}_r^2\right)^{-1} +$$

$$\frac{(\bar{Y} - \bar{X})^2}{4(S_x^2 + S_y^2 - 2\hat{\delta}_r^2)^3} \times \left[\frac{2\hat{\delta}_r^4}{n_x - 1} + \frac{2\hat{\delta}_r^4}{n_y - 1} + \frac{8\hat{\delta}_r^2}{n_f}\right].$$

(see appendix 1). Therefore, an approximate $(1 - \alpha) \times 100$ percent confidence interval for $\delta$ is given by

$$\hat{\delta} \pm z_{\alpha/2} \sqrt{\text{V\text{\textbar}r(}\hat{\delta})},$$

where $z_{\alpha/2}$ is the $1 - \frac{\alpha}{2}$ standard normal percentile, with the corresponding interval for $A$ being

$$\left\{\Phi(\hat{\delta} - z_{\alpha/2} \sqrt{\text{V\text{\textbar}r(}\hat{\delta})}), \Phi(\hat{\delta} + z_{\alpha/2} \sqrt{\text{V\text{\textbar}r(}\hat{\delta})})\right\}.$$
For $n_X = n_Y = 50$ and $n_f = 19$, the observed coverage of the proposed confidence interval is close to its nominal value for $R > 0.5$, even though the nominal value sometimes lies outside the binomial confidence interval. However, for small $R$, the coverage tends to be conservative for $A < 0.7$ and low for $A \geq 0.7$. With a larger sample size ($n_X = n_Y = 100$, $n_f = 49$), the estimated confidence interval shows some improvement, with observed coverages being reasonably close to their nominal values for all $A$ with $R > 0.2$, although, in some cases, the target value is outside the binomial confidence interval. When we consider the situation of an unequal number of cases and controls, dealing with an extreme case of $n_X = 900$ and $n_Y = 50$, the results are very similar to those previously presented. Interestingly, when $n_X$, $n_Y$, and $n_f$ are large ($n_X = n_Y = 1,000$, $n_f = 199$), the coverage probability is very good for almost every situation except $R = 0.2$, which indicates that a better coverage is obtained as a function of sample size. We have also increased the sample size to be $n_X = n_Y = 10,000$, $n_f = 1,999$ and found that the observed coverage was very close to the nominal value, even for $R = 0.2$. Although these results indicate that asymptotically the coverage is correct, for practical purposes such large sample sizes are not available, and thus, we conclude that for $R = 0.2$ our confidence interval does not performed satisfactory. However, $R = 0.2$ is an extreme case. Note that for $\hat{\sigma}^2 = \sigma^2 = 2$, $R = 0.2$ implies $\hat{\sigma}^2 = 4\hat{\sigma}^2$, which rarely occurs in practice (11).

All of the above indicate that for a wide range of sample sizes and parameter values, equation 2 provides an effective confidence interval for the area under the ROC curve.

We have computed the percentage of negative variance estimators. As one might expect, this percentage is very much

### Table 1. Observed coverage probabilities of confidence intervals

<table>
<thead>
<tr>
<th>$A$</th>
<th>$R$</th>
<th>$n_X = 50$, $n_Y = 50$, $n_f = 19$</th>
<th>$n_X = 100$, $n_Y = 100$, $n_f = 49$</th>
<th>$n_X = 900$, $n_Y = 50$, $n_f = 199$</th>
<th>$n_X = 1,000$, $n_Y = 1,000$, $n_f = 199$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>1.0</td>
<td>3.0</td>
<td>0.5</td>
</tr>
<tr>
<td>0.6</td>
<td>0.2</td>
<td>97.8</td>
<td>98.3</td>
<td>97.7</td>
<td>97.2</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2</td>
<td>97.4</td>
<td>97.2</td>
<td>97.2</td>
<td>96.8</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2</td>
<td>95.4</td>
<td>95.5</td>
<td>95.2</td>
<td>95.3</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>93.3</td>
<td>93.6</td>
<td>93.4</td>
<td>93.5</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2</td>
<td>95.9</td>
<td>95.8</td>
<td>95.5</td>
<td>96.2</td>
</tr>
<tr>
<td>0.6</td>
<td>0.2</td>
<td>96.4</td>
<td>96.3</td>
<td>96.5</td>
<td>95.5</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2</td>
<td>95.2</td>
<td>95.1</td>
<td>94.8</td>
<td>95.1</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2</td>
<td>89.3</td>
<td>88.7</td>
<td>89.2</td>
<td>91.3</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2</td>
<td>94.4</td>
<td>94.1</td>
<td>93.9</td>
<td>93.5</td>
</tr>
<tr>
<td>0.6</td>
<td>0.2</td>
<td>96.1</td>
<td>96.5</td>
<td>96.3</td>
<td>95.9</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2</td>
<td>95.6</td>
<td>95.7</td>
<td>96.0</td>
<td>95.3</td>
</tr>
<tr>
<td>0.9</td>
<td>0.2</td>
<td>84.9</td>
<td>85.9</td>
<td>84.9</td>
<td>88.9</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2</td>
<td>92.9</td>
<td>92.7</td>
<td>92.3</td>
<td>93.6</td>
</tr>
<tr>
<td>0.6</td>
<td>0.2</td>
<td>95.5</td>
<td>95.5</td>
<td>95.6</td>
<td>96.1</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2</td>
<td>96.2</td>
<td>96.9</td>
<td>95.9</td>
<td>95.9</td>
</tr>
</tbody>
</table>

*Values in italics are the estimated coverage probabilities whose 95% confidence intervals (based on a binomial sample of 7,300 simulated data sets) do not include the targeted coverage probabilities.*
affected by sample size. In the case of small sample sizes, the percentage was as high as 25 percent for \( R = 0.2 \), about 10 percent for \( R = 0.4 \), 2.5 percent for \( R = 0.6 \), and less than 0.5 percent for \( R = 0.8 \). These results were similar for various areas under the curve. As sample sizes increase, the percentage of negative variance estimates decreases, and for very large sample sizes, it becomes negligible. Finally, we have evaluated the bias of \( \hat{A} \) and found that it was, at most, 2 percent and in many cases much less. Since the bias was so small, for the sake of brevity, we omitted presenting the individual biases.

**EXAMPLE**

Oxygen free radicals are believed to be related to the processes of aging and chronic diseases, as well as to atherosclerotic coronary heart disease (17). Recent studies have suggested that oxidative modification of low density lipoproteins may be a critical factor (17) in the atherosclerotic process.

A thiobarbituric acid reaction substance (TBARS) is a biomarker that measures subproducts of lipid peroxidation and has been proposed as a discriminating measurement of cardiovascular disease cases and healthy controls.

A population-based sample of randomly selected residents of Erie and Niagara counties, New York, who were aged 35–79 years is the focus of this investigation (18). The New York State Department of Motor Vehicles driver’s license rolls were used as the sampling frame for adults between ages 35 and 64 years, while the elderly sample (ages 65–79) was randomly selected from the Health Care Financing Administration.

Blood samples, physical measurements, and a detailed questionnaire on various behavioral and physiologic patterns were obtained from study participants. Owing to cost and logistical considerations, sample replications were not obtained from the participants.

After the exclusion of participants with a history of cancer (60 subjects) and/or incomplete information on TBARS (68 subjects) and of non-White participants (75 subjects), a total of 474 White men and 494 White women were selected for subjects) and of non-White participants (75 subjects), a total of 60 subjects) and/or incomplete information on TBARS (68 subjects) and logistical considerations, sample replications were not

A thiobarbituric acid reaction substance (TBARS) is a biomarker that measures subproducts of lipid peroxidation and has been proposed as a discriminating measurement of cardiovascular disease cases and healthy controls.

A population-based sample of randomly selected residents of Erie and Niagara counties, New York, who were aged 35–79 years is the focus of this investigation (18). The New York State Department of Motor Vehicles driver’s license rolls were used as the sampling frame for adults between ages 35 and 64 years, while the elderly sample (ages 65–79) was randomly selected from the Health Care Financing Administration.

Blood samples, physical measurements, and a detailed questionnaire on various behavioral and physiologic patterns were obtained from study participants. Owing to cost and logistical considerations, sample replications were not obtained from the participants.

After the exclusion of participants with a history of cancer (60 subjects) and/or incomplete information on TBARS (68 subjects) and of non-White participants (75 subjects), a total of 474 White men and 494 White women were selected for subjects) and of non-White participants (75 subjects), a total of 60 subjects) and/or incomplete information on TBARS (68 subjects) and logistical considerations, sample replications were not obtained from the participants.

For the reliability study, we obtain the following:

\[
\hat{A} = 0.642 \\
R = 0.59
\]

where \( \hat{A} \) is obtained by substituting estimates for the parameters in the formula for \( R \). The values of \( \hat{n}_x, n_y, \) and \( n_x \) in this particular study are similar to those used for the simulations in table 1, column 3. Furthermore, since table 1 indicates that the confidence intervals computed according to equation 2 have good coverage properties for \( R > 0.2 \), we can expect them to operate reasonably well in our example with \( R = 0.59 \).

Not correcting for measurement error results in \( \hat{A} = 0.642 \) and in 0.550, 0.725 as the unadjusted 95 percent confidence interval for \( A \). Applying the methods of the section Statistical Inference in the Presence of Random Measurement Error, we obtain the adjusted area estimate to be \( \hat{A}_c = 0.735 \), with the corresponding corrected 95 confidence interval being 0.600, 0.888.

Correction for measurement error increased the estimate of the area under the ROC curve and shifted the confidence interval to include much higher values. Use of the uncorrected results understates the effectiveness of TBARS as a biomarker capable of discriminating between subjects with and those without cardiovascular disease.

**DISCUSSION**

In this paper, we have shown that an approximate confidence interval for the area under the ROC curve, which takes measurement error into account through a reliability study, can be obtained by using the Delta method. Our simulation study indicates that this method has good coverage properties for a wide range of situations. If measurement error is not taken into account in the construction of the confidence interval, the effectiveness of the biomarker can be seriously understated. We assumed equality of the measurement errors for both the controls and the cases. If this assumption is not appropriate, reliability studies on both cases and controls should be conducted to allow separate estimation of the measurement errors. Applying the same methods as in the section Statistical Inference in the Presence of Random Measurement error, we now obtain

\[
\text{Var}(\hat{\delta}) = \left[ \frac{S_x^2}{n_x} + \frac{S_y^2}{n_y} \right] \times \left( \frac{S_x^2 - \sigma_x^2 + S_y^2 - \sigma_y^2}{n_x} \right)^{-1} + (\hat{x} - \overline{y})^2 \times \frac{4(S_x^2 - \sigma_x^2 + S_y^2 - \sigma_y^2)}{4(S_x^2 - \sigma_x^2 + S_y^2 - \sigma_y^2)} \times
\]

where \( \hat{\delta} \) is the transformed data yield the following results:

\[
\overline{x} = 0.604 \\
\overline{y} = 0.450 \\
n_x = 928 \\
n_y = 40 \\
S_x^2 = 0.0913 \\
S_y^2 = 0.0886
\]

A reliability study to estimate the variance due to random measurement error in the analysis of TBARS was conducted on a convenience sample of 10 participants. Twelve-hour fasting blood samples were obtained for seven women and three men over a period of 6 months to estimate the measurement error variability. The blood samples were obtained every month on the same day of each female’s menstrual cycle and every month on the same calendar day for each male. From the reliability study, we obtain the following:

\[
\hat{\sigma}_x^2 = 0.0567 \\
n_y = 41 \\
R = 0.59
\]
where $\hat{\sigma}_1^2$, $n_f$, and $\hat{\sigma}_2^2$, $n_t$ are the estimated variances due to random measurement error and their corresponding degrees of freedom for the controls and cases of the reliability studies, respectively. Applying equation 3 to equation 2 results in the approximate confidence interval for this case.

The use of correction methods for random measurement error remains controversial in epidemiologic studies. These methods require the assumption that the variance due to random measurement error is known or can be reliably estimated. The lack of good estimates of this variance can limit the usefulness of the correction. When the reliability or validation study involves only small numbers of subjects and replications, the estimated variance of measurement error will have a large variability, leading to uncertainty with respect to corrections. Moreover, if the measurement error varies over subgroups of the population (risk status, gender, etc.) and these subgroups are not taken into account by the reliability study, the estimates of measurement error variance may be seriously biased, leading to bias in the adjustment.

We evaluated the sensitivity of our method to deviations from distributional assumptions. We found that the method is susceptible to departure from the normality. When distributional assumptions were grossly violated, our method underestimated the area under the ROC curve, and the corresponding confidence intervals provided coverage that is less than its nominal value. In cases in which nonnormality is suspected, suitable transformations can often be used (19).

We have varied our choice of small numbers to replace the negative variance estimates. We looked at seven different alternatives. When $S_2^2 < \hat{\sigma}_1^2 < S_1^2 < \hat{\sigma}_2^2$, $S_1^2(S_2^2)$ was replaced in the simulation with $1.4 \times \hat{\sigma}_2^2$, $1.2 \times \hat{\sigma}_2^2$, $1.1 \times \hat{\sigma}_2^2$, $1.01 \times \hat{\sigma}_2^2$, $1.005 \times \hat{\sigma}_2^2$, $1.001 \times \hat{\sigma}_2^2$, or we replaced $\hat{\sigma}_2^2$ with $0.8 \times \min(S_1^2, S_2^2)$, we found the coverage probability to be robust for the seven different options. Since $1.001 \times \hat{\sigma}_2^2$ provided slightly better results than the other choices did, this option was used for the results reported in table 1. The clinical implications of this method are important for early detection and prevention of diseases. If one is using the corrected area results in a marker being considered a good discriminator, while the uncorrected area indicates poor discrimination, then to use this marker in future subjects it is necessary to reduce the measurement error. If the source of the error is technical, for example, laboratory equipment, several approaches to reduce random measurement error are available. An average over replications can be taken. Alternatively, the laboratory quality control protocol may be revised and improved to minimize different sources of unwanted variability (e.g., to shorten elapsed time between when the sample was taken and when it was analyzed). Sometimes, new and more effective measuring instruments and methods can be used. However, if the source of error is intraindividual variability, a standardized protocol should be implemented to reduce this source of variation.

In conclusion, we showed that measurement error could result in a serious understatement of the effectiveness of a biomarker. A point estimate and a confidence interval procedure for the area under the ROC curve, which corrects for random measurement error, were presented. This procedure assumes the availability of data from a reliability study of the biomarker.

REFERENCES

APPENDIX 1

In the measurement error case, \( x_i \) and \( y_i \) are independent normal random variables with means and variances \( \mu_x, \sigma_x^2 + \sigma_e^2 \) and \( \mu_y, \sigma_y^2 + \sigma_e^2 \). Consequently, \( \bar{x} \) and \( \bar{y} \) are independent normal random variables with means and variances \( \mu_x, (\sigma_x^2 + \sigma_e^2)/n_x \) and \( \mu_y, (\sigma_y^2 + \sigma_e^2)/n_y \), respectively.

Thus,

\[
\hat{\mu} = \bar{y} - \bar{x} \sim \text{Normal} \left( \mu_y - \mu_x, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y} \right) \tag{A1}
\]

and

\[
(n_x - 1) \frac{S_x^2}{\sigma_x^2 + \sigma_e^2} \sim \chi^2_{n_x - 1}, \quad (n_y - 1) \frac{S_y^2}{\sigma_y^2 + \sigma_e^2} \sim \chi^2_{n_y - 1}, \quad (n_y) \frac{\hat{\sigma}_e^2}{\sigma_e^2} \sim \chi^2_{n_y}.
\]

are all mutually independent.

In addition,

\[
\hat{\sigma}^2 = \hat{\sigma}_x^2 + \hat{\sigma}_y^2 = (S_x^2 - \hat{\sigma}_e^2) + (S_y^2 - \hat{\sigma}_e^2)
\]

so that

\[
\hat{\delta} = \frac{\hat{\mu}}{\hat{\sigma}}. \tag{A2}
\]

We use the Delta method (13) to obtain the approximate variance of \( \hat{\delta} \), noting the independence of \( \hat{\mu} \) and \( \hat{\sigma} \).

\[
\text{Var}(\hat{\delta}) = \left( \frac{\partial \hat{\delta}}{\partial \hat{\mu}} \right)^2 \text{Var}(\hat{\mu}) + \left( \frac{\partial \hat{\delta}}{\partial \hat{\sigma}} \right)^2 \text{Var}(\hat{\sigma}) = \frac{1}{\sigma^2} \times \text{Var}(\hat{\mu}) + \frac{\mu^2}{\sigma^4} \times \text{Var}(\hat{\sigma}). \tag{A3}
\]

\( \text{Var}(\hat{\mu}) \) is given in equation A1, while for \( \text{Var}(\hat{\sigma}) \), we use the Delta method again to obtain

\[
\text{Var}(\hat{\sigma}) = \left( \frac{\partial (\hat{\sigma}^2)}{\partial \sigma^2} \right)^2 \times \text{Var}(\hat{\sigma}^2) = \frac{1}{4 \times \sigma^2} \times \text{Var}(\hat{\sigma}^2)
\]

\[
= \frac{1}{4 \sigma^2} \times \text{Var}(S_x^2 + S_y^2 - 2\hat{\sigma}_e^2) = \frac{1}{4 \sigma^2} \times \left[ \text{Var}(S_x^2) + \text{Var}(S_y^2) + 4 \times \text{Var}(\hat{\sigma}_e^2) \right]
\]

\[
= \frac{1}{4 \sigma^2} \times \left[ \frac{2(\sigma_x^2 + \sigma_e^2)^2}{n_x - 1} + \frac{2(\sigma_y^2 + \sigma_e^2)^2}{n_y - 1} + \frac{8(\sigma_e^2)^2}{n_f} \right]. \tag{A4}
\]

Applying equations A1 and A4 to equation A3 results in

\[
\text{Var}(\hat{\delta}) = \frac{1}{\sigma^2} \times \left[ \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y} + \frac{\sigma_e^2}{n_f} \right] + \frac{\mu^2}{4 \times \sigma^3} \times \left[ \frac{2 \times (\sigma_x^2 + \sigma_e^2)^2}{n_x - 1} + \frac{2 \times (\sigma_y^2 + \sigma_e^2)^2}{n_y - 1} + \frac{8(\sigma_e^2)^2}{n_f} \right] \tag{A5}
\]

Substituting the estimates for the unknown parameters in equation A5 results in

\[
\text{Var}(\hat{\delta}) = \left[ \frac{S_x^2}{n_x} + \frac{S_y^2}{n_y} \right] \times (S_x^2 - \hat{\sigma}_e^2 + S_y^2 - \hat{\sigma}_e^2)^{-1} + \frac{(\bar{x} - \bar{y})^2}{4(S_x^2 - \hat{\sigma}_e^2 + S_y^2 - \hat{\sigma}_e^2)} \times \left[ \frac{2S_x^4}{n_x - 1} + \frac{2S_y^4}{n_y - 1} + \frac{8\hat{\sigma}_e^4}{n_f} \right].
\]