Published in final edited form as:

Stat Med. 2013 November 30; 32(27): . doi:10.1002/sim.5878.

Adjustment for Measurement Error in Evaluating Diagnostic Biomarkers by Using an Internal Reliability Sample

Matthew T. White and Sharon X. Xie^{*,†}

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

Summary

Biomarkers are often measured with error due to imperfect lab conditions or temporal variability within subjects. Using an internal reliability sample of the biomarker, we develop a parametric bias-correction approach for estimating a variety of diagnostic performance measures including sensitivity, specificity, the Youden index with its associated optimal cut-point, positive and negative predictive values, and positive and negative diagnostic likelihood ratios when the biomarker is subject to measurement error. We derive the asymptotic properties of the proposed likelihood-based estimators and show that they are consistent and asymptotically normally distributed. We propose confidence intervals for these estimators and confidence bands for the receiver operating characteristic curve. We demonstrate through extensive simulations that the proposed approach removes the bias due to measurement error and outperforms the naïve approach (which ignores the measurement error) in both point and interval estimation. We also derive the asymptotic bias of naïve estimates and discuss conditions in which naïve estimates of the diagnostic measures are biased towards estimates produced when the biomarker is ineffective (i.e., when sensitivity equals 1 – specificity) or are anti-conservatively biased. The proposed method has broad biomedical applications and is illustrated using a biomarker study in Alzheimer's disease. We recommend collecting an internal reliability sample during the biomarker discovery phase in order to adequately evaluate the performance of biomarkers with careful adjustment for measurement error.

Keywords

biomarkers; diagnostic measures; maximum likelihood; measurement error; replicate data

1. Introduction

As biomarkers have grown in importance in epidemiological studies, drug discovery, translational research, clinical trials, and the exciting new era of “personalized medicine,” it is critical to develop analytical approaches to reliably evaluate their performance. Many biomarkers are obtained with measurement error due to imperfect lab conditions (e.g., operator error, contamination, variable storage conditions) or temporal variability within subjects. In various application areas, such as studying the use of biomarkers in plasma or cerebrospinal fluid (CSF) to differentiate between subjects with and without Alzheimer's disease (AD), there is evidence that the measurement error variance may be large relative to the variance of the “true” biomarker [1]. Under these circumstances, it is important to

^{*}Correspondence to: Sharon X. Xie, Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Drive, Philadelphia, Pennsylvania 19104, USA.

[†]sxie@mail.med.upenn.edu

examine the reasonableness of measurement model assumptions and to employ analytic methods that behave well under extreme measurement error.

In this paper, we study the effectiveness of a continuous biomarker in differentiating diseased from non-diseased individuals. Standard diagnostic measures are often used to evaluate the performance of biomarkers. These measures include sensitivity, specificity, the Youden index with its associated optimal cut-point, the receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), positive predictive value (PPV), negative predictive value (NPV), positive diagnostic likelihood ratio (PLR), and negative diagnostic likelihood ratio (NLR). The biomarker, however, is usually measured with error. Let X_{ik} denote the k^{th} replicate obtained on the i^{th} subject in the diseased group. Also, let Y_{jl} be the l^{th} replicate obtained on the j^{th} subject in the non-diseased group. We assume X_{ik} and Y_{jl} follow a classical measurement error model

$$\begin{aligned} X_{ik} &= x_i + \varepsilon_{ik}, \quad i=1, \dots, n_1, \quad k=1, \dots, m_{1i}, \\ Y_{jl} &= y_j + \eta_{jl}, \quad j=1, \dots, n_0, \quad l=1, \dots, m_{0j}, \end{aligned} \quad (1)$$

where $n_1(n_0)$ is the number of subjects in the diseased (non-diseased) group, and $m_{1i}(m_{0j})$ is the number of replicate observations for subject $i(j)$. These replicate data form an internal reliability sample. We assume that the true values of the biomarker for diseased (non-diseased) subjects, $x_i(y_j)$, are independent and normally distributed with mean $\mu_1(\mu_0)$ and variance $\sigma_1^2(\sigma_0^2)$. We also assume that the errors $\varepsilon_{ik}(\eta_{jl})$ are independent and normally distributed with mean 0 and variance $\tau_1^2(\tau_0^2)$ and further assume that $x_i, y_j, \varepsilon_{ik}$, and η_{jl} are independent. To simplify notation, we sometimes use the subscript D to specify that the parameter refers to group D ($D=1$ for the diseased group and $D=0$ for the non-diseased group).

There is growing belief that for biomedical studies, diagnostic data should be analyzed with careful adjustment for the effect of measurement error [2–7]. Most proposals to date focus on AUC estimation in the presence of measurement error. Coffin and Sukhatme [2] and Faraggi [8] showed that not taking measurement error into account underestimates the AUC and could give the erroneous impression that a biomarker is not effective. Thus, appropriate adjustment for measurement error is important in order to minimize the chance of missing biomarkers that are potentially useful for disease treatment and prevention. Coffin and Sukhatme [2, 3] developed a bias correction method for the point estimator of the AUC in the presence of measurement error using parametric and non-parametric approaches. Schisterman *et al.* [4] presented both an estimator and a confidence interval (CI) for the AUC when an external reliability study is performed to evaluate measurement error. Li *et al.* [9] presented interval estimation of the AUC using the method of variance estimates recovery. Reiser [10] proposed an estimator and its CI of the AUC by adjusting the measurement error using an internal reliability sample through the maximum likelihood estimation procedures described by Searle *et al.* [11]. Reiser [10] also briefly discussed how to estimate the ROC curve itself but did not consider confidence bands for the curve. Tosteson *et al.* [12] studied the effect of measurement error on the ROC curve and proposed a bias-corrected estimator of the ROC curve with its pointwise confidence band when the measurement error variance is heterogeneous and the distribution of the measurement error is non-normal. Vexler *et al.* [13] developed methods to account for both measurement error and pooling design in the context of ROC curve estimation. Perkins *et al.* [6] developed ROC curve inference for a biomarker subject to a limit of detection and measurement error. Long *et al.* [7] studied robust statistical methods for the AUC and other diagnostic measures when biomarkers are subject to batch/experiment-specific measurement errors. Perkins and Schisterman [5] studied the bias of the naïve Youden index estimate and the bias-corrected

Youden index estimate when an external validation study is available, and Ruopp *et al.* [14] created bias-corrected estimates of the Youden index and optimal cut-point when observations are subject to a lower limit of detection.

There are three new contributions to the literature from this paper which we summarize below. First, we study the effect of measurement error by deriving the analytical expressions of the asymptotic biases of naïve sensitivity, specificity, PPV, NPV, PLR, and NLR estimates (Section 2). Second, we propose bias-corrected estimators for sensitivity, specificity, the Youden index with its associated optimal cut-point, PPV, NPV, PLR, and NLR in the presence of internal reliability data for an error-prone biomarker using a maximum likelihood approach described by Searle *et al.* [11] that was used by Reiser [10] for AUC estimation (Section 3). Asymptotic distribution theory for these new estimators is given along with asymptotic variance estimators (Section 4). We also propose confidence intervals for these diagnostic measures (Section 4), constructed using appropriate transformations. Third, we propose methods to construct the rectangular confidence region for a point on the ROC curve as well as pointwise and simultaneous confidence bands for the ROC curve by using an internal reliability sample when the measurement error variance is unknown (Section 4).

The rest of the article is organized as follows. The performance of the proposed bias-corrected estimates in comparison to the naïve estimates in small to moderate sample sizes are studied through extensive simulations (Section 5). The simulations consider different magnitudes of measurement error variance and proportions of subjects with replicate data. This is followed by an illustrative example of a biomarker study (Section 6) and discussion (Section 7). The discussion points to applications in which evaluating noisy biomarkers is particularly useful.

2. Asymptotic Bias in Naïve Diagnostic Measures

By convention, larger values of a biomarker are associated with disease (i.e., $\mu_1 > \mu_0$). For each potential cut-point c of the continuous biomarker, an individual is classified as diseased if $x_i < c$ or $y_j < c$ or classified as non-diseased if $x_i > c$ or $y_j > c$. For a specific cut-point c ,

true sensitivity (Se) is equal to $1 - \Phi\left(\frac{c - \mu_1}{\sigma_1}\right)$ under the normality assumption specified in measurement error model (1), where $\Phi(\cdot)$ denotes the cumulative distribution function (cdf)

of a $\mathcal{N}(0, 1)$ random variable. Similarly, true specificity (Sp) is equal to $\Phi\left(\frac{c - \mu_0}{\sigma_0}\right)$ for a specific cut-point c . The ROC curve is a plot of sensitivity versus $1 - \text{specificity}$ for all cut-points. Naïve estimators are based on the average of M_1 (M_0) randomly selected observations per subject, denoted as X_i ($\bar{Y}_j$) in the diseased (non-diseased) group, where $M_1 = \min\{m_{11}, \dots, m_{1n1}\}$ and $M_0 = \min\{m_{01}, \dots, m_{0n0}\}$. When the number of replicates differs among diseased (non-diseased) subjects, we do not average over all replicates per subject since the X_i ($\bar{Y}_j$) would no longer be identically distributed. The Naïve sensitivity estimator

$$\widehat{Se}_{nai} = 1 - \Phi\left(\frac{c - \hat{\mu}_{1,nai}}{\hat{\sigma}_{1,nai}}\right) \text{ where } \hat{\mu}_{1,nai} = \frac{1}{n_1} \sum_{i=1}^{n_1} \bar{X}_i \text{ and } \hat{\sigma}_{1,nai} = \sqrt{\frac{\sum_{i=1}^{n_1} (\bar{X}_i - \bar{\mu}_{1,nai})^2}{n_1 - 1}}. \text{ The naïve specificity estimator is } \widehat{Sp}_{nai} = \Phi\left(\frac{c - \hat{\mu}_{0,nai}}{\hat{\sigma}_{0,nai}}\right),$$

where $\hat{\mu}_{0,nai} = \frac{1}{n_0} \sum_{j=1}^{n_0} \bar{Y}_j$ and $\hat{\sigma}_{0,nai} = \sqrt{\frac{\sum_{j=1}^{n_0} (\bar{Y}_j - \hat{\mu}_{0,nai})^2}{n_0 - 1}}$. Tosteson *et al.* [12] presented a plot of ROC curves for different magnitudes of measurement error using naïve

sensitivity and specificity but did not explicitly derive the expression of bias for the naïve sensitivity and specificity. In this section, we derive the formula of the bias of Naïve sensitivity and specificity estimates as a function of measurement error. Using basic probability arguments, as the sample size in each group goes to infinity, the naïve sensitivity

estimator can be shown to converge in probability to $Se_{nai} = 1 - \Phi\left(\frac{\Phi^{-1}(1 - Se)}{\sqrt{1 + \tau_1^2/(M_1\sigma_1^2)}}\right)$, and the naïve specificity estimator can be shown to converge in probability to

$Sp_{nai} = \Phi\left(\frac{\Phi^{-1}(Sp)}{\sqrt{1 + \tau_0^2/(M_0\sigma_0^2)}}\right)$, where $\Phi^{-1}(\cdot)$ denotes the inverse of the cdf of a $\mathcal{N}(0, 1)$ random variable. Therefore, the asymptotic bias in these naïve estimates can be expressed as

$$Se_{nai} - Se = 1 - \Phi\left(\frac{\Phi^{-1}(1 - Se)}{\sqrt{1 + \tau_1^2/(M_1\sigma_1^2)}}\right) - Se \quad \text{for sensitivity and}$$

$$Sp_{nai} - Sp = \Phi\left(\frac{\Phi^{-1}(Sp)}{\sqrt{1 + \tau_0^2/(M_0\sigma_0^2)}}\right) - Sp \quad \text{for specificity. Additionally, we derived an expression for the exact bias in finite samples. Using the independence of the sample mean and variance for normal random variables, the exact bias can be derived as}$$

$$E(\widehat{Se}_{nai}) - Se = 1 - \Phi\left(\frac{\Phi^{-1}(1 - Se)}{\sqrt{1 + \tau_1^2/(M_1\sigma_1^2)}} \times \frac{\sqrt{\frac{n_1-1}{2}} \Gamma\left(\frac{n_1-2}{2}\right)}{\Gamma\left(\frac{n_1-1}{2}\right)}\right) - Se \quad \text{for sensitivity and}$$

$$E(\widehat{Sp}_{nai}) - Sp = \Phi\left(\frac{\Phi^{-1}(Sp)}{\sqrt{1 + \tau_0^2/(M_0\sigma_0^2)}} \times \frac{\sqrt{\frac{n_0-1}{2}} \Gamma\left(\frac{n_0-2}{2}\right)}{\Gamma\left(\frac{n_0-1}{2}\right)}\right) - Sp \quad \text{for specificity, where}$$

$\Gamma(\cdot)$ is the Gamma function. As the sample size in each group goes to infinity, the limit of the exact bias for sensitivity and specificity is the same as the asymptotic bias. The asymptotic bias for naïve sensitivity (specificity) depends upon the true diagnostic parameter value Se (Sp), the minimum number of replicates M_1 (M_0), and the noise to signal ratio $\theta_1^2 = \tau_1^2/\sigma_1^2$ ($\theta_0^2 = \tau_0^2/\sigma_0^2$). This index was also used in Faraggi [8] and Reiser [10].

Schisterman *et al.* [4] utilized a related “reliability index” $R = (\sigma_1^2 + \sigma_0^2)/(\sigma_1^2 + \sigma_0^2 + \tau_1^2 + \tau_0^2)$, where values close to 0 (1) correspond to a large (small) noise to signal ratio. Indeed, if we write the overall noise to signal ratio as $\theta^2 = (\tau_1^2 + \tau_0^2)/(\sigma_1^2 + \sigma_0^2)$, we see that $\theta^2 = 1/R - 1$. The naïve sensitivity estimator is asymptotically unbiased when the true sensitivity is 0, 0.5, or 1. A true value of 0 or 1 corresponds to degenerate cases where the mean or cut-point is infinite or the biomarker has 0 variance. A true sensitivity of 0.5 indicates that the biomarker performs equivalently to a coin toss in identifying diseased subjects. In this case, the biomarker is just noise, so adding more noise in the form of measurement error will not affect the accuracy. If $\theta_1^2/M_1 > 0$, the asymptotic bias of $\widehat{Se}_{nai}$ is positive when $0 < Se < 0.5$ and negative when $0.5 < Se < 1$, with the magnitude of the bias increasing to $|0.5 - Se|$ as $\theta_1^2/M_1 \rightarrow \infty$. Thus, the naïve sensitivity estimate is biased towards the case in which the biomarker performs similarly to a coin toss. The bias of $\widehat{Sp}_{nai}$ follows a similar pattern. In this paper, we classify a biomarker as ineffective when $Se = 1 - Sp$ (corresponding to a 45-degree line for the ROC curve) and as effective when $Se < 1 - Sp$. Thus, the ROC curve based on naïve estimates of sensitivity and specificity is biased towards the 45-degree line when the biomarker is effective.

Diagnostic likelihood ratios are functions of sensitivity and specificity used to evaluate the prognostic efficacy of a biomarker. They quantify the change in the odds of disease after

observing the result of the biomarker. Replacing Se with $\widehat{Se}_{nai}$ and Sp with $\widehat{Sp}_{nai}$ in the standard definitions for PLR and NLR, the asymptotic bias in the naïve estimates can be

expressed as $\frac{Se_{nai}}{1 - Sp_{nai}} - \frac{Se}{1 - Sp}$ for PLR and $\frac{1 - Se_{nai}}{Sp_{nai}} - \frac{1 - Se}{Sp}$ for NLR. For an ineffective biomarker, $PLR = NLR = 1$, indicating that the biomarker result does not influence a patient's post-test odds of disease. The top two plots in Figure 1 display the asymptotic bias of naïve PLR and NLR estimates as a function of the true parameter value.

When $\theta_1^2/M_1 = \theta_0^2/M_0$, the asymptotic biases in naïve estimates of PLR and NLR are always conservative (i.e., biased towards 1) when the biomarker is effective and unbiased when the biomarker is ineffective. When $\theta_1^2/M_1 \neq \theta_0^2/M_0$ and the biomarker is effective, the bias in naïve estimates of PLR and NLR may sometimes be anti-conservative, as can be seen in Figure 1 (e.g., positive bias in the solid curve when the true PLR is between 1 and 1.5). However, we can show through simple calculations of the above bias expressions that if the true sensitivity and specificity are both greater than 0.5, naïve estimates of PLR and NLR will always be conservatively biased, regardless of the values of θ_1^2/M_1 and θ_0^2/M_0 .

Predictive values are functions of sensitivity, specificity, and disease prevalence () and are used to determine how well a biomarker predicts disease status. Replacing Se with $\widehat{Se}_{nai}$ and Sp with $\widehat{Sp}_{nai}$ in the standard definitions for PPV and NPV, the asymptotic bias in the naïve

PPV estimate can be expressed as $\left(1 + \frac{1 - Sp_{nai}}{Se_{nai}} \times \frac{1 - \rho}{\rho}\right)^{-1} - \text{true PPV}$, and the

asymptotic bias in the naïve NPV estimate can be expressed as $\left(1 + \frac{1 - Se_{nai}}{Sp_{nai}} \times \frac{\rho}{1 - \rho}\right)^{-1} - \text{true NPV}$. For an ineffective biomarker, $PPV =$ and $NPV = 1 -$, indicating that the biomarker is not informative in determining disease status. The bottom two plots in Figure 1 show the asymptotic bias of the naïve PPV and NPV estimates as a function of the true parameter value when the disease prevalence = 0.5. Like PLR and NLR, when

$\theta_1^2/M_1 = \theta_0^2/M_0$, the asymptotic biases in naïve PPV and NPV estimates are conservative when the biomarker is effective and unbiased when the biomarker is ineffective. When $\theta_1^2/M_1 \neq \theta_0^2/M_0$ and the biomarker is effective, the bias in naïve estimates of PPV and NPV may sometimes be anti-conservative, as can be seen in Figure 1 (e.g., positive bias away from = 0.5 in the solid curve when the true PPV is between 0.5 and 0.6). However, if the true sensitivity and specificity are both greater than 0.5, naïve estimates of PPV and NPV will always be conservatively biased, regardless of the values of θ_1^2/M_1 and θ_0^2/M_0 . The bias pattern is similar for other values of disease prevalence (see Web-based Supporting Figure 1). The analytic expressions for the biases of the naïve sensitivity, specificity, PLR, NLR, PPV, and NPV estimates have not been presented previously in the literature.

Youden [15] proposed a measure of overall diagnostic effectiveness, J , defined as the maximum of sensitivity + specificity – 1 over all possible cut-points c . Perkins and Schisterman [5] showed that the true optimum cut-point where J occurs is

$$c^* = \begin{cases} \frac{\mu_0(r^2-1)-e+r\sqrt{e^2+(r^2-1)\sigma_0^2\log(r^2)}}{r^2-1} & \text{if } r \neq 1 \\ \frac{\mu_1+\mu_0}{2} & \text{if } r=1, \end{cases} \quad (2)$$

where $e = \mu_1 - \mu_0$ and $r = \sigma_1 / \sigma_0$. The true value of J is thus $\Phi\left(\frac{\mu_1 - c^*}{\sigma_1}\right) + \Phi\left(\frac{c^* - \mu_0}{\sigma_0}\right) - 1$. The naïve estimators for c^* and J are obtained by replacing the true unknown mean and standard deviation with their naïve estimates described previously. The biases of the naïve estimates of c^* and J were studied previously in Perkins and Schisterman [5].

3. Bias-corrected Method to Adjust for Measurement Error

Note that the observed replicates are correlated within each subject. Thus, we use a maximum likelihood estimation procedure for the 1-way random-effects model (equation (1)) described by Searle *et al.* [11] (page 86), which was used by Reiser [10] for AUC estimation. This procedure produces maximum likelihood estimates (MLEs) of μ_D , σ_D^2 , and τ_D^2 (denoted $\hat{\mu}_D$, $\hat{\sigma}_D^2$, and $\hat{\tau}_D^2$, respectively) which we use to obtain consistent and asymptotically normally distributed estimators of sensitivity, specificity, the Youden index with its associated optimal cut-point, PLR, NLR, PPV, and NPV. We require the following

notation: $N_1 = \sum_{i=1}^{n_1} m_{1i}$, $\bar{X}_{i\cdot} = \sum_{k=1}^{m_{1i}} X_{ik} / m_{1i}$, and $\lambda_i = \tau_1^2 + m_{1i} \sigma_1^2$. We obtain the MLEs $\hat{\mu}_1$, $\hat{\sigma}_1^2$, and $\hat{\tau}_1^2$ by maximizing the log-likelihood equation

$$l = -\frac{1}{2} N_1 \log(2\pi) - \frac{1}{2} (N_1 - n_1) \log(\tau_1^2) - \frac{1}{2} \sum_{i=1}^{n_1} \log \lambda_i - \frac{1}{2\tau_1^2} \sum_{i=1}^{n_1} \sum_{k=1}^{m_{1i}} (X_{ik} - \mu_1)^2 + \frac{1}{2\tau_1^2} \sum_{i=1}^{n_1} \frac{\sigma_1^2 m_{1i} (\bar{X}_{i\cdot} - \mu_1)^2}{\lambda_i}. \quad (3)$$

The MLEs $\hat{\mu}_0$, $\hat{\sigma}_0^2$, and $\hat{\tau}_0^2$ are obtained similarly for the non-diseased group.

Under the normality assumption for the biomarker and using the invariance property of

MLEs, the MLEs of sensitivity and specificity can be written as $\widehat{Se} = 1 - \Phi\left(\frac{c - \hat{\mu}_1}{\hat{\sigma}_1}\right)$ and $\widehat{Sp} = \Phi\left(\frac{c - \hat{\mu}_0}{\hat{\sigma}_0}\right)$, respectively.

MLEs of PLR, NLR, PPV, NPV, the Youden index, and the optimal cut-point (denoted $\widehat{PLR}$, $\widehat{NLR}$, $\widehat{PPV}$, $\widehat{NPV}$, $\widehat{J}$, and $\widehat{c}^*$, respectively) are obtained by replacing the true unknown parameters with the corresponding MLEs in their respective definitions.

4. Asymptotic Properties and Proposed Confidence Intervals of the Bias-corrected Estimators

All asymptotic properties refer to when the total number of subjects n_D in group D ($D = 1, 0$). We also assume n_1/n_0 converges to a bounded number as n_1 and n_0 . We let $\xrightarrow{P}$ stand for convergence in probability. Derivation of the asymptotic properties is provided in Web Appendix A. As MLEs, $\hat{\mu}_D$, $\hat{\sigma}_D^2$, and $\hat{\tau}_D^2$ are consistent and asymptotically normally distributed. Using the asymptotic results for MLEs for the 1-way random-effects model in Searle *et al.* [11], the asymptotic covariance estimator of $\hat{\mu}_1$, $\hat{\sigma}_1^2$, and $\hat{\tau}_1^2$ is

$$\widehat{\text{var}} \begin{bmatrix} \hat{\mu}_1 \\ \hat{\tau}_1^2 \\ \hat{\sigma}_1^2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n_1} \frac{m_{1i}}{\hat{\lambda}_i} & 0 & 0 \\ 0 & \frac{N_1 - n_1}{2\hat{\tau}_1^4} + \frac{1}{2} \sum_{i=1}^{n_1} \frac{1}{\hat{\lambda}_i^2} & \frac{1}{2} \sum_{i=1}^{n_1} \frac{m_{1i}}{\hat{\lambda}_i^2} \\ 0 & \frac{1}{2} \sum_{i=1}^{n_1} \frac{m_{1i}}{\hat{\lambda}_i^2} & \frac{1}{2} \sum_{i=1}^{n_1} \frac{m_{1i}^2}{\hat{\lambda}_i^2} \end{bmatrix}^{-1},$$

Where $\hat{\lambda}_i = \hat{\tau}_1^2 + m_{1i} \hat{\sigma}_1^2$. The asymptotic covariance estimator of $\hat{\mu}_0$, $\hat{\sigma}_0^2$, and $\hat{\tau}_0^2$, denoted as $\widehat{\text{var}}(\hat{\mu}_0, \hat{\tau}_0^2, \hat{\sigma}_0^2)'$, is obtained similarly for the non-diseased group.

As functions of $\hat{\mu}_D$ and $\hat{\sigma}_D^2$, $\widehat{Se} \xrightarrow{P} Se$ and $\widehat{Sp} \xrightarrow{P} Sp$ by the continuous mapping theorem. Using the asymptotic normality property of $\hat{\mu}_D$ and $\hat{\sigma}_D^2$, and the delta method, we can show that $\sqrt{n_1}(\widehat{Se} - Se)$ converges in distribution to a mean 0 normal variate (see Web Appendix A). The asymptotic variance estimator for $\widehat{Se}$ is

$$\widehat{\text{var}}(\widehat{Se}) = \frac{1}{2\pi\hat{\sigma}_1^2} \exp \left\{ - \left(\frac{c - \hat{\mu}_1}{\hat{\sigma}_1} \right)^2 \right\} \left\{ \widehat{\text{var}}(\hat{\mu}_1) + \left(\frac{c - \hat{\mu}_1}{2\hat{\sigma}_1^2} \right)^2 \widehat{\text{var}}(\hat{\sigma}_1^2) \right\}. \quad (4)$$

Results follow similarly for specificity.

Since sensitivity and specificity are restricted to be between 0 and 1, we recommend creating confidence intervals on the Φ^{-1} transformed scale and back-transforming the endpoints to the original scale. The variances on the transformed scale are derived using the delta method. Thus, the 100(1 - α)% CI for Se is

$$\Phi \left(\Phi^{-1}(\widehat{Se}) \pm \Phi^{-1}(1 - \alpha/2) \sqrt{\frac{\widehat{\text{var}}(\hat{\mu}_1)}{\hat{\sigma}_1^2} + \frac{(c - \hat{\mu}_1)^2 \widehat{\text{var}}(\hat{\sigma}_1^2)}{4\hat{\sigma}_1^6}} \right). \text{ The CI for } Sp \text{ is similar,}$$

substituting $\widehat{Sp}$ for $\widehat{Se}$ and using parameter estimates from the non-diseased group.

The ROC curve is a plot of sensitivity versus 1 - specificity for all cut-points c . The estimated ROC curve is a plot of $\widehat{Se}$ versus $1 - \widehat{Sp}$ for all cut-points c . At each cut-point (corresponding to a point on the ROC curve), we can construct a rectangular 100(1 - α)% confidence region for $(1 - Sp, Se)$ through the above transformed CIs. Using the results in Pepe [16], if $((1 - \widehat{Sp})_L, (1 - \widehat{Sp})_U)$ and $(\widehat{Se}_L, \widehat{Se}_U)$ are 100(1 - α)% univariate confidence intervals for $(1 - Sp)$ and Se , where $\alpha = 1 - (1 - \alpha)^{1/2}$, then the 100(1 - α)% rectangular confidence region for $(1 - Sp, Se)$ is

$$R \equiv \left((1 - \widehat{Sp})_L, (1 - \widehat{Sp})_U \right) \times \left(\widehat{Se}_L, \widehat{Se}_U \right). \quad (5)$$

Instead of creating a confidence region around a single point on the ROC curve, we can develop a confidence band around the entire ROC curve. Pepe [16] shows that for each value t of 1 - specificity, the binormal ROC curve can be expressed as $\text{ROC}(t) = \Phi(a + b\Phi^{-1}(t))$, where $a = (\mu_1 - \mu_0)/\sigma$ and $b = \sigma/\sigma_1$. Thus $\text{ROC}(t)$ can be

estimated by $\widehat{ROC}(t) = \Phi(\hat{a} + \hat{b}\Phi^{-1}(t))$, where $\hat{a} = (\mu_1^* - \mu_0^*)/\sigma_1^*$ and $\hat{b} = \sigma_0^*/\sigma_1^*$. The variance estimator of $\widehat{ROC}(t)$, denoted $\widehat{var}(\widehat{ROC}(t))$, is provided in Web Appendix B.

Since the ROC curve can be expressed as a known function () of a straight line, we can use a Working-Hotelling-type band for a straight line as described in Ma and Hall [17] to obtain a simultaneous confidence band. Thus, we propose a $100(1 - \alpha)\%$ simultaneous confidence band for $ROC(t)$ on an interval $[0,1]$ as

$$\Phi\left(\hat{a} + \hat{b}\Phi^{-1}(t) \pm k\sqrt{\widehat{var}(\widehat{ROC}(t))}\right), \quad (6)$$

Where $k = \sqrt{-2\log(\alpha)}$. For a fixed t , a $100(1 - \alpha)\%$ pointwise confidence band for $ROC(t)$ can be constructed using $k = z_{1-\alpha/2}$.

PLR, NLR, PPV, and NPV are functions of sensitivity and specificity. Using the consistency and asymptotic normality of sensitivity and specificity, we can show that $\widehat{PLR}$, $\widehat{NLR}$, $\widehat{PPV}$, and $\widehat{NPV}$ are consistent and asymptotically normally distributed (see Web Appendix A). Since the ranges of PLR, NLR, PPV, and NPV are restricted, we recommend creating confidence intervals on appropriately transformed scales and back-transforming the endpoints to the original scale (see Web Appendix C). Finally, consistency and asymptotic normality of $\hat{e}$ and $\hat{f}$ can be established using arguments similar to those used for sensitivity. Variance estimators and appropriate confidence intervals for these measures are described in Web Appendix C.

5. Simulation Study

Our method to estimate and adjust for measurement error using an internal reliability sample is motivated by the fact that measurement error biases estimates of the true biomarker variance in the diseased and non-diseased groups (σ_1^2 and σ_0^2 , respectively) and thus biases sensitivity, specificity, the ROC curve, the Youden index and associated optimal cut-point, PLR, NLR, PPV, and NPV, away from their true values. A series of simulation studies was conducted to assess the ability of our method to account for different magnitudes of measurement error variance in a sample where 35%, 60%, or 100% of subjects in the diseased and non-diseased groups had four replicates ($m_{1i} = m_{0j} = 4$). The properties assessed include the estimated bias (estimator–true parameter), observed sample standard deviation, estimated standard error ($\widehat{se}$), and coverage rate of the nominal 95% CI based on appropriate transformations described in Section 4. Naïve estimators were computed as described in Section 2, and their confidence intervals were based on the same type of transformations used in the proposed estimators.

Simulations were conducted for sample sizes of $n_1 = n_0 = 50, 100$, and 150 in each group. Due to space limitations, Tables 1, 3, and 4 present simulations with the sample size of 150 in each group. The corresponding simulations with sample sizes of 50 and 100 in each group are presented in Web-based Supporting Tables 1, 2, and 3. Table 2 presents simulations of sample sizes of 50, 100, and 150 in each group with 60% having replicate data for the three types of confidence intervals or bands for the true ROC curve. All simulations assume a disease prevalence of 0.5, true sensitivity Se of 0.85, and true specificity Sp of 0.70. The corresponding true PLR, NLR, PPV, and NPV are 2.83, 0.21, 0.74, and 0.82, respectively. The true biomarker was simulated as $\mathcal{N}(6.5, 0.3)$ in the diseased group and as $\mathcal{N}(5.7, 0.2)$ in the non-diseased group. These simulation parameters were of similar magnitudes to those

observed in the real data example in Section 6. The true ROC curve was generated by $(1.465 + 0.816 t^{-1}(t))$, $t \in [0, 1]$. The optimal cut-point is therefore 6.12, corresponding to $J = 0.58$. Measurement error variances of $\tau_D^2 = 0.1$ and $\tau_D^2 = 0.5$ were examined, corresponding to an overall moderate ($\tau^2 = 0.4$) or large ($\tau^2 = 2$) noise to signal ratio, respectively. The moderate noise to signal ratio value is similar to that observed in our real data example.

Results are based on 10,000 simulation runs for each combination of τ_D^2 and fraction of subjects with replicate data. We note that for all simulations presented in Tables 1, 3, and 4, the observed sample standard deviation was found to correspond adequately with the average of the standard deviation estimates (mean $\hat{s}_{\hat{\theta}}$) from the asymptotic theory for both the naïve and bias-corrected estimates.

Table 1 compares the performance of the naïve and bias-corrected estimates of sensitivity and specificity. We see that at each value of τ_D^2 the naïve estimate is negatively biased for both sensitivity and specificity, with the magnitude of bias increasing as τ_D^2 increases. This agrees with our analytical predictions from Section 2. In contrast, the proposed estimates (described in Section 3) exhibit very little bias even when the noise to signal ratio is large and the size of the reliability sample is small. As a result, the CI (based on the t^{-1} transformation) coverage rates for the bias-corrected estimates remain close to the nominal 95% level while the coverage rates decrease dramatically for the naïve estimates as τ_D^2 increases.

Table 2 compares the coverage of the naïve and bias-corrected estimates of the 95% simultaneous confidence band for the true ROC curve, the 95% pointwise confidence band for the true ROC curve at the true value of $1 - \text{specificity}$ (0.3), and the 95% rectangular confidence region around the true value (0.3, 0.85) on the true ROC curve under various sample sizes. We see that the proposed estimates perform well with coverage rates close to the nominal 95% level, especially when the sample size increases. In contrast, the coverage rates decrease dramatically for the naïve estimates as τ_D^2 increases. Similar results were found for the cases where 35% and 100% of subjects had replicate data (data not shown).

Tables 3 and 4 present the corresponding results from the sensitivity and specificity estimates for the naïve and bias-corrected estimates of PLR and NLR and of PPV and NPV, respectively (see Section 3). We see that at each value of τ_D^2 , the naïve estimate is negatively biased for PLR, PPV, and NPV and positively biased for NLR, with the magnitude of the bias increasing as τ_D^2 increases. This agrees with our analytical predictions from Section 2 (see Figure 1). In contrast, the proposed estimates exhibit very little bias even when the noise to signal ratio is large and the size of the reliability sample is small. As a result, the CI (based on appropriate transformations described in Web Appendix C) coverage rates for the proposed estimates remain close to the nominal 95% level while the coverage rates decrease dramatically for the naïve estimates as τ_D^2 increases.

The bias pattern for PPV and NPV is similar for other values of disease prevalence (see Web-based Supporting Table 4). While the standard errors for the proposed PPV and NPV estimators are non-linear functions of the population prevalence (see Web Appendix A and Web-based Supporting Table 4), we see that the CI coverage rates of PPV and NPV are not affected by disease prevalence. This is due to the relationship between the logit of PPV (NPV) and the log of PLR (NLR) (see Web Appendix C), and it can be shown that the CI coverage rate for PPV (NPV) is the same as the CI coverage rate for PLR (NLR) (see Tables 3 and 4). The performances of the remaining diagnostic measures remain unaffected for other values of disease prevalence since they do not depend upon disease prevalence.

We also assessed the performance of the naïve and bias-corrected estimates of the optimal cut-point and Youden's index J in simulations. For simulations in which 35% of 150 subjects in each group had replicate data, there is only a slight positive bias of 0.012 ($\tau_D^2=0.1$) and 0.028 ($\tau_D^2=0.5$) in the naïve estimates of the optimal cut-point, resulting in CI coverage rates of the naïve estimates close to the nominal 95% level (0.965 and 0.950, respectively). This is due to the fact that $\sigma_1^2 \approx \sigma_0^2$ and $\tau_1^2 \approx \tau_0^2$, and thus the optimal cut-point mainly depends upon the means of the biomarker values in each group, as can be seen in equation (2). The bias-corrected estimates of the optimal cut-point exhibit very little bias (-0.001 and -0.004 for $\tau_D^2=0.1$ and $\tau_D^2=0.5$, respectively), resulting in CI coverage rates close to the nominal 95% level (0.948 and 0.954, respectively). For $\tau_D^2=0.1$ and $\tau_D^2=0.5$ the naïve estimates are negatively biased for J (-0.077 and -0.224, respectively) while the proposed estimates exhibit very little bias (0.003 and 0.012, respectively). As a result, as τ_D^2 increases from 0.1 to 0.5, the CI coverage rates decrease dramatically for the naïve estimates of J (0.513 and 0, respectively) while CI coverage rates for the proposed estimates (based on the log(-log) transformation) remain close to the nominal 95% level (0.954 and 0.964, respectively). Similar results were found for the cases where 60% and 100% of subjects had replicate data and for the sample size of 100 in each group (data not shown). When the sample size is only 50 in each group, the coverage probability for the bias-corrected estimator for J is slightly higher (around 97%) than the nominal 95% level.

The simulations presented in Tables 1, 3, and 4 were repeated for sample sizes of 50 and 100 in each group. The results are presented in Web-based Supporting Tables 1, 2, and 3. The relative performance of the naïve and bias-corrected estimates in these simulations was similar to that presented above for a sample size of 150 in each group except for PLR. For the large measurement error case ($\tau_D^2=0.5$), when the sample size is 50 and the percent of replicates is 35% or 60%, the bias-corrected estimates for PLR display a similar magnitude of bias as the naïve estimates. This is because the small sample size and small fraction of subjects with replicate data yield highly variable estimates of specificity, and a specificity estimate near 1 results in a large estimate for PLR since the denominator, $(1 - \text{specificity})$, is near 0. As the size of each group increases to above 100 or the size of the reliability sample increases to 100%, the proposed PLR estimate becomes more accurate while the naïve estimate remains biased. This is expected since the proposed estimator is consistent.

In many settings, there may be more error variance in the diseased group than the non-diseased group. Additional simulations were conducted in which the overall noise to signal ratio τ_1^2/τ_0^2 was kept at the levels presented above (0.4 or 2), but $\tau_1^2=2\tau_0^2$. The results are presented in Web-based Supporting Tables 5–8. When $\tau_1^2=2\tau_0^2$, the proposed bias-corrected estimators still exhibited very little bias with appropriate coverage, but the standard error estimates were slightly larger (smaller) for $\widehat{Se}$ ($\widehat{Sp}$), as compared to the case in which $\tau_1^2=\tau_0^2$, due to the relatively larger (smaller) error variance in the diseased (non-diseased) group. The standard error estimates for other bias-corrected diagnostic measures also changed. In addition, the magnitude of the bias increased (decreased) in the naïve sensitivity (specificity) estimators, resulting in worse (better) coverage as compared to the case in which $\tau_1^2=\tau_0^2$. Interestingly, the biases for the PPV (NPV) and PLR (NLR) parameters were smaller (larger). This is due to the fact that the biases for these measures are non-linear functions of the given values for sensitivity and specificity as well as the magnitude of the measurement error in each group; different values for these parameters may exhibit very different bias patterns for the likelihood ratios and predictive values. Finally, when $\tau_1^2=2\tau_0^2$, the measurement error differentially affects the naïve estimates of the true biomarker variances

in each group, making them appear more dissimilar. As a result, the magnitude of the bias in the naïve cut-point estimate increased and the coverage decreased as the measurement error variance increased. The bias-corrected estimator for the optimal cut-point was not similarly affected.

6. Tau as a Biomarker for Alzheimer's Disease - An Example

We illustrate our methods by considering data from an AD biomarker study conducted at the University of Pennsylvania (Penn) Alzheimer's Disease Core Center which was designed to evaluate the relationship between tau protein and AD. Previous studies have shown tau protein to be elevated in AD, forming tangled masses in the brain [18]. Because only palliative treatments are available after AD onset, it is of great importance to identify biomarkers to better classify and treat this multifactorial syndrome. In the Penn AD biomarker study, CSF was obtained by lumbar puncture, and the level of CSF tau was measured by ELISA [19, 20]. A CSF tau sample was taken once for each of 418 subjects, and the assay was performed independently of the clinical diagnosis. A total of 363 subjects (87%) were clinically diagnosed as having AD, and the remaining 55 subjects (13%) were non-diseased individuals. As discussed in the introduction, there is evidence that CSF biomarkers can be noisy. Thus, investigators in this study randomly selected a portion of subjects to have their tau samples analyzed more than once in the lab to form an internal reliability sample. Subjects' tau samples were analyzed up to seven times, with 44% and 84% of AD and non-AD subjects, respectively, having replicate data. The observed values of tau exhibited a large positive-skew and were log-transformed for analysis. After transformation, the observed data appeared to be normally distributed. The tau measurement unit is pg/mL. We will omit the unit in the presentation below since the results are presented on a log-scale. Naïve estimators were computed as described in Section 2.

Due to the invasiveness of the lumbar puncture procedure, each subject had only one CSF sample taken. The CSF sample for each subject was analyzed once for subjects not in the reliability sample and was analyzed more than once for subjects in the reliability sample. Therefore, we can estimate the measurement error due to imperfect lab conditions but cannot estimate the temporal variability of the true biomarker value within subjects. To ensure the validity of our 1-way random effects model (equation (1)), we need to verify the additive error assumption and the normality and independence assumptions for the error terms. The additive error assumption was assessed for each group by plotting the intra-individual standard deviation against the mean [21]. The plots (not shown) did not show any structure, indicating that the log-transformation was reasonable and that the additive error assumption was appropriate. For each group, analyses of model residuals were used to assess the joint normality and independence assumptions for the error terms. Normal quantile-quantile plots of the residuals (not shown) were approximately linear, Mardia's tests for multivariate skewness and kurtosis [22] were not significant, and pairwise correlations among the residuals did not differ significantly from 0. Thus, the joint normality and independence assumptions were not violated in our data.

Using the MLE estimators that maximize the log-likelihood (equation (3) in Section 3), the estimated mean and variance of the true biomarker in the AD group are 6.28 and 0.30, respectively. In the non-diseased group, the estimated mean and variance of the true biomarker are 5.56 and 0.21, respectively. The estimated measurement error variances in the AD and non-diseased groups are 0.10 and 0.07, respectively. Figure 2 (top) presents the estimated ROC curves using naïve and proposed MLE estimates of sensitivity and specificity across cut-points of the biomarker.

We see that the naïve ROC curve is biased towards the 45-degree line (and thus the naïve AUC is biased towards 0.5), which agrees with results shown previously in the literature [2]. In contrast, the bias-corrected ROC curve is closer to the left and upper borders of the positive unit quadrant. Using the bias-corrected estimator for the ROC curve, we constructed a 95% simultaneous confidence band for the true ROC curve (expression 6 in Section 4). With 95% confidence, the true ROC curve lies completely within the confidence band. Using results in Sections 3 and 4, the bias-corrected optimal cut-point $\hat{c}$ was found to be 5.95 (95% CI: 5.59–6.31) with $\hat{f} = 0.54$ (95% CI: 0.43–0.64), corresponding to a sensitivity of 0.73 (95% CI: 0.69–0.77) and a specificity of 0.81 (95% CI: 0.70–0.89). Using the rectangular confidence region formula (equation (5) in Section 4), the 95% confidence rectangle based on this point $(1 - 0.81, 0.73)$ on the bias-corrected ROC curve is $(0.10, 0.32) \times (0.68, 0.78)$. Thus with 95% confidence, the true $1 - \text{specificity}$ is between 0.10 and 0.32, and the true sensitivity is between 0.68 and 0.78. The CI for $1 - \text{specificity}$ is wider than the CI for sensitivity since there are fewer subjects in the non-diseased group. The naïve estimate of the optimal cut-point is 5.96 (95% CI: 5.49–6.26), with a naïve estimate of Youden's index of 0.48 (95% CI: 0.36–0.58), corresponding to a naïve sensitivity of 0.69 (95% CI: 0.65–0.73) and a naïve specificity of 0.78 (95% CI: 0.67–0.87). The naïve optimal cut-point is close to the bias-corrected optimal cutpoint because the estimated variances of the true biomarker as well as the estimated variances of the measurement error in both the diseased and non-diseased groups are similar. The value of the naïve AUC is 0.81 (95% CI: 0.75–0.87), and the bias-corrected AUC is 0.85 (95% CI: 0.79–0.90 using the CI formula in Reiser [10]). Thus, tau protein appears to be an effective biomarker for classifying AD status.

The bottom left plot in Figure 2 presents the estimated PLR as a function of NLR across all cut-points of the biomarker. At each cut-point, we see that the naïve estimate of PLR is lower than the bias-corrected estimate of PLR. Also, we see that at each cut-point, the naïve estimate of NLR is higher than the bias-corrected estimate of NLR. The point where $\text{PLR} = \text{NLR} = 1$ corresponds to the case in which the biomarker is ineffective (i.e. $Se = 1 - Sp$). We see that at all cut-points, the naïve estimates of PLR and NLR are biased towards the point corresponding to an ineffective biomarker. At the bias-corrected optimal cut-point, the bias-corrected PLR and NLR estimates are 3.76 (95% CI: 2.30–6.15) and 0.33 (95% CI: 0.27–0.41), respectively. In comparison, at the naïve optimal cut-point, the naïve PLR and NLR estimates are 3.21 (95% CI: 1.99–5.17) and 0.39 (95% CI: 0.33–0.47). Pepe [16] states that the diagnostic likelihood ratio can be thought of as the ratio of post-test odds of being diseased to pre-test odds of being diseased. The prevalence of AD among individuals who were eligible for the study was quite high (0.72). Prior to conducting the test, the odds of having AD correspond to the ratio of disease prevalence to $(1 - \text{disease prevalence})$, $/(1 - 0.72) = 0.72/0.28 = 2.57$, in the study population. With a positive (negative) test result, the post-test odds of AD equal the pre-test odds of AD multiplied by the PLR (NLR). Using the bias-corrected optimal cut-point, with a positive test result, the post-test odds of AD are increased from 2.57 to $2.57 \times 3.76 = 9.66$. With a negative test result, the post-test odds of AD are decreased to $2.57 \times 0.33 = 0.85$. These results provide further evidence that tau protein appears to be an effective biomarker for AD.

The bottom right plot of Figure 2 presents the estimated PPV as a function of NPV across all cut-points of the biomarker. We see that for each cut-point, the naïve estimates of PPV and NPV are lower than the bias-corrected estimates of PPV and NPV. The prevalence of AD was 0.72, and this value was used in the estimation of PPV and NPV. The point where $\text{PPV} = \text{NPV} = 1 - \text{prevalence}$ corresponds to the case in which the biomarker is completely ineffective. Once again, we see that at all cut-points, the naïve estimates of PPV and NPV are biased towards the point corresponding to an ineffective biomarker. At the bias-corrected optimum cut-point, the bias-corrected PPV and NPV estimates are 0.91 (95% CI: 0.86–0.94)

and 0.54 (95% CI: 0.49–0.59), respectively. In comparison, at the naïve optimal cut-point, the naïve PPV and NPV estimates are 0.89 (95% CI: 0.84–0.93) and 0.50 (95% CI: 0.44–0.54). From the bias-corrected estimates, we see that 91% of subjects with a positive test result have AD while 46% of subjects with a negative test result have AD. Thus, CSF tau appears to be useful in clinical practice in the current study population.

7. Discussion

We have studied the effect of measurement error on naïve estimates of various diagnostic performance measures through theoretical derivations, simulations, and a data analysis example. We found that ignoring measurement error may lead to serious bias of diagnostic performance measures, and we therefore proposed parametric, bias-corrected methods that yield consistent and asymptotically normally distributed estimators of diagnostic measures by using an internal reliability sample. In particular, we showed that for biomarkers whose true sensitivity and specificity are not 0, 0.5, or 1, the naïve estimates of sensitivity and specificity are always conservatively biased towards 0.5, the point at which the biomarker performs equivalently to a coin toss. We also found that the biases in naïve estimates of other diagnostic performance measures, including PLR, NLR, PPV, and NPV, are always conservative (towards the point at which $Se = 1 - Sp$) when the true sensitivity and specificity are both above 0.5 but can be anticonservative in other settings. As the minimum number of replicates M_D in each group increases, the magnitude of the bias in naïve estimates decreases. In simulations (Section 5), naïve estimates had coverage rates close to the nominal 95% level when all subjects had 4 replicate observations and there was a moderate amount of measurement error. However, when only a portion of subjects have replicate data or when measurement error is relatively large, naïve estimates perform poorly. The proposed estimators perform well under all study conditions.

Our approach is parametric in the sense that both the biomarker and measurement error are assumed to follow a normal distribution. In our real data example, the observed biomarker values exhibited a large positive skew. After using a log transformation, the data appeared to be normally distributed. It is therefore important to conduct the model-checking procedures outlined in Section 6 to ensure the validity of the bias-corrected estimates. The assumptions were not violated for our data example, but there may be cases in which these assumptions may not be met post-transformation. To address this issue, additional non-parametric methods to relax the normality assumption for the biomarker and measurement error are needed.

In contrast with using an external validation study to quantify the magnitude of measurement error, an internal reliability study can avoid strong assumptions about the homogenous characteristics of the main study and the external validation study [21]. In engineering and computer science applications, the overall probability of misclassification is sometimes used as another measure of diagnostic performance. The bias-corrected estimator for overall probability of misclassification can be derived using similar techniques for bias-corrected PPV or NPV estimators. We note that Li *et al.* [9] presented an interesting method of variance estimates recovery (MOVER) to develop confidence intervals when the sampling distribution of the statistic is asymmetric. Since our estimators are asymptotically normal, the MOVER approach leads to identical results as the delta method in the context of this paper. We have confirmed this using our data example (results not shown). The proposed approach is not only applicable to biomarker evaluation but also to any diagnostic test evaluation.

The proposed method does not require that every subject has replicate data. In fact, due to cost and logistical reasons, many reliability studies can only be done on a subsample of the

cohort, in which case the above results will apply with $m_{1j} = m_{0j} = 1$ except in the subsample. Interesting study design issues arise with regard to the size of the reliability subsample that is needed to adequately accommodate measurement error. For example, in epidemiological studies, if the total budget for assessing biomarkers is fixed, it is valuable to determine the optimum size of the study cohort, optimum size of the reliability subsample, and optimum size of the number of replicates for each subject within the reliability subsample. These design issues are currently under investigation.

If a noisy biomarker is found effective using the proposed method, it can help us in several ways to advance the scientific process. First, it reduces the likelihood of missing important biomarkers that could potentially be beneficial in developing therapeutic interventions. For example, a recent study shows that insulin polypeptides are associated with classifying depression in older adults [23], suggesting that diabetes drugs may be effective in the treatment of depression. In this setting, biomarkers are not used to diagnose disease but to identify effective new drugs. Therefore, it is plausible to design clinical trials to examine the effectiveness of diabetes drugs on depression treatment. Second, if the noisy biomarker will be used for disease diagnosis in future subjects, then the next step is to reduce the measurement error. If the source of the error is due to imperfect lab conditions, the laboratory quality control procedures should be improved and more precise lab techniques may need to be developed. If the source of the error is due to temporal variability within subjects, data collection procedures should be standardized. In summary, we recommend collecting an internal reliability subsample during the biomarker discovery phase in order to adequately evaluate the performance of biomarkers with careful adjustment for measurement error.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported by NIH grants including 5 T32 DK07785 (Renal and Urologic Biostatistics Training Grant), AG-10124 (the Alzheimer's Disease Core Center), AG-32953, AG-17586, and NS-053488. We are grateful to all of our colleagues in the University of Pennsylvania Center for Neurodegenerative Disease Research and the University of Pennsylvania Memory Center. The authors thank Dr. Justine Shults for her critical review of the original version of the paper as well as two anonymous reviewers for their excellent suggestions.

References

1. Mattsson N, Blennow K, Zetterberg H. Inter-laboratory variation in cerebrospinal fluid biomarkers for Alzheimer's disease: united we stand, divided we fall. *Clinical Chemistry and Laboratory Medicine*. 2010; 48:603–607.10.1515/CCLM.2010.131 [PubMed: 20201744]
2. Coffin M, Sukhatme S. Receiver operating characteristic studies and measurement errors. *Biometrics*. 1997; 53:823–837. [PubMed: 9333348]
3. Coffin, M.; Sukhatme, S. A parametric approach to measurement errors in receiver operating characteristic studies. In: Jewell, NP.; Kimber, AC.; Ting Lee, ML.; Whitmore, GA., editors. *Lifetime Data: Models in Reliability and Survival Analysis*. Kluwer, Boston: 1995. p. 71-75.
4. Schisterman EF, Faraggi D, Reiser B, Trevisan M. Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. *American Journal of Epidemiology*. 2001; 154:174–179.10.1093/aje/154.2.174 [PubMed: 11447052]
5. Perkins NJ, Schisterman EF. The Youden index and optimal cut-point corrected for measurement error. *Biometrical Journal*. 2005; 47:428–441.10.1002/bimj.200710415 [PubMed: 16161802]
6. Perkins NJ, Schisterman EF, Vexler A. Generalized ROC curve inference for a biomarker subject to a limit of detection and measurement error. *Statistics in Medicine*. 2009; 28:1841–1860.10.1002/sim.3575 [PubMed: 19340817]

7. Long Q, Flanders WD, Fedirko V, Bostick RM. Robust statistical methods for analysis of biomarkers measured with batch/experiment-specific errors. *Statistics in Medicine*. 2010; 29:361–370.10.1002/sim.3796 [PubMed: 20020422]
8. Faraggi D. The effect of random measurement error on receiver operating characteristic (ROC) curves. *Statistics in Medicine*. 2000; 19:61–70.10.1002/(SICI)1097-0258(20000115)19:1<61::AID-SIM297>3.0.CO;2-A [PubMed: 10623913]
9. Li Y, Koval JJ, Donner A, Zuo GY. Interval estimation for the area under the receiver operating characteristic curve when data are subject to error. *Statistics in Medicine*. 2010; 29:2521–2531.10.1002/sim.4015 [PubMed: 20684006]
10. Reiser B. Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. *Statistics in Medicine*. 2000; 19:2115–2129.10.1002/1097-0258(20000830)19:16<2115::AID-SIM529>3.0.CO;2-M [PubMed: 10931515]
11. Searle, SR.; Casella, G.; McCulloch, CE. *Variance Components*. Wiley; New York: p. 1992
12. Tosteson TD, Buonaccorsi JP, Demidenko E, Wells WA. Measurement error and confidence intervals for ROC curves. *Biometrical Journal*. 2005; 47:409–416.10.1002/bimj.200310159 [PubMed: 16161800]
13. Vexler A, Schisterman EF, Liu A. Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. *Statistics in Medicine*. 2008; 27:280–296.10.1002/sim.3035 [PubMed: 17721905]
14. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal*. 2008;419–430. textbf50. 10.1002/bimj.200710415 [PubMed: 18435502]
15. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950; 3:32–35.10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3 [PubMed: 15405679]
16. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; Oxford; 2003.
17. Ma G, Hall WJ. Confidence bands for receiver operating characteristic curves. *Medical Decision Making*. 1993; 13:191–197. [PubMed: 8412547]
18. Arnold SE, Hyman BT, Flory J, Damasio AR, Van Hoesen GW. The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with Alzheimer's disease. *Cerebral Cortex*. 1991; 1:103–116.10.1093/cercor/1.1.103 [PubMed: 1822725]
19. Vandermeeren M, Mercken M, Vanmechelen E, Six J, Van de Voorde A, Martin JJ, Cras P. Detection of proteins in normal and Alzheimer's disease cerebrospinal fluid with a sensitive sandwich enzyme-linked immunosorbent assay. *Journal of Neurochemistry*. 1993; 61:1828–1834.10.1111/j.1471-4159.1993.tb09823.x [PubMed: 8228996]
20. Clark CM, Xie SX, Chittams J, Ewbank D, Peskind E, Galasko D, Morris JC, McKeel DW Jr, Farlow M, Weitlauf SL, Quinn J, Kaye J, Knopman D, Arai H, Doody RS, DeCarli C, Leight S, Lee VY, Trojanowski JQ. CSF tau and -amyloid: How well do these biomarkers reflect autopsy confirmed dementia diagnoses? *Archives of Neurology*. 2003; 60:1696–1702.10.1001/archneur.60.12.1696 [PubMed: 14676043]
21. Carroll, RJ.; Ruppert, D.; Stefanski, LA. *Measurement Error in Nonlinear Models*. Chapman & Hall/CRC; Boca Raton: p. 1995
22. Mardia KV. Measures of multivariate skewness and kurtosis with applications. *Biometrika*. 1970; 57:519–530.10.1093/biomet/57.3.519
23. Arnold SE, Xie SX, Leung YY, Wang LS, Kling MA, Han X, Kim EJ, Wolk DA, Bennett DA, Chen-Plotkin A, Grossman M, Hu W, Lee VMY, Scott Mackin R, Trojanowski JQ, Wilson RS, Shaw LM. Plasma biomarkers of depressive symptoms in older adults. *Translational Psychiatry*. 2012; 2:e65.10.1038/tp.2011.6326 [PubMed: 22832727]

Figure 1.

Asymptotic bias of naïve estimators (based on an average of $M_0 = M_1 = 4$ replicates per subject) for positive diagnostic likelihood ratio (PLR), negative diagnostic likelihood ratio (NLR), positive predictive value (PPV) and negative predictive value (NPV) as a function of the true parameter value, with specificity $p = 0.7$, disease prevalence $\pi = 0.5$, and noise to signal ratio θ_D^2 in the diseased (healthy) group of 4 (0.5) (—), 2 (2) (---), or 0.5 (4) (.....).

Figure 2.

Top: ROC curve for $\log(\text{Tau})$. Lines are based on naïve (—) and bias-corrected (-----) estimates in the Alzheimer's disease biomarker study. Lighter dashed lines represent the 95% simultaneous confidence band for the ROC curve using the bias-corrected estimator.

Bottom left: Positive likelihood ratio (PLR) as a function of negative likelihood ratio (NLR). Lines are based on naïve (—) and bias-corrected (-----) estimates. The single point (•) corresponds to the case where the biomarker is ineffective, i.e., sensitivity = 1 – specificity.

Bottom right: Positive predictive value (PPV) as a function of negative predictive value (NPV). Lines are based on naïve (—) and bias-corrected (-----) estimates. The single point (•) corresponds to the case where the biomarker is ineffective, i.e., sensitivity = 1 – specificity.

MLEs in diseased group: mean of true $\log(\text{Tau}) = 6.28$, variance of true $\log(\text{Tau}) \hat{\sigma}_1^2 = 0.30$, measurement error variance $\hat{\tau}_1^2 = 0.10$. MLEs in nondiseased group: $\hat{\mu}_0 = 5.56$, $\hat{\sigma}_0^2 = 0.21$, $\hat{\tau}_0^2 = 0.07$. The total sample size $N = 418$, and the disease prevalence = 0.72

Simulation Results for Sensitivity and Specificity. Note: Results are based on 10,000 runs with 35%, 60%, or 100% of subjects having replicate (4) observations. The coverage (nominal 95% level) is based on the τ_D^2 transformed confidence intervals described in Section 4. Naive estimates are based on M observations per subject in each group. The true biomarker is normal (6.5, 0.3) in diseased subjects and normal (5.7, 0.2) in non-diseased subjects. The measurement error is normal (0, τ_D^2) in each group D. The true sensitivity Se = 0.85, specificity Sp = 0.7, and disease prevalence = 0.5. The sample size in each group is 150.

Table 1

Parameter (True Value)	Percent Replicates	M	Method	Bias			Standard Deviation			Mean($\hat{Se}$)			Percent Coverage		
				$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$	$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$	$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$	$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$
Sensitivity (0.85)	35	1	Naive Proposed	-0.035 0.001	-0.113 0.003	0.025 0.027	0.029 0.040	0.025 0.039	0.025 0.027	0.025 0.027	0.028 0.039	68.2 94.8	1.0 96.2		
	60	1	Naive Proposed	-0.034 0.001	-0.113 0.002	0.025 0.026	0.028 0.035	0.028 0.034	0.025 0.025	0.025 0.025	0.028 0.034	68.7 94.9	0.9 95.6		
	100	4	Naive Proposed	-0.010 0.001	-0.042 0.001	0.023 0.024	0.025 0.029	0.025 0.029	0.024 0.024	0.024 0.024	0.025 0.029	93.1 95.5	57.6 95.5		
Specificity (0.70)	35	1	Naive Proposed	-0.034 0.001	-0.090 0.006	0.032 0.036	0.033 0.057	0.035 0.055	0.036 0.037	0.036 0.037	0.035 0.055	86.8 95.3	24.3 96.4		
	60	1	Naive Proposed	-0.034 0.001	-0.090 0.003	0.032 0.036	0.033 0.050	0.035 0.049	0.036 0.036	0.036 0.036	0.035 0.049	86.3 94.8	23.4 96.0		
	100	4	Naive Proposed	-0.010 0.001	-0.040 0.002	0.031 0.033	0.032 0.042	0.036 0.042	0.037 0.033	0.037 0.033	0.036 0.042	97.0 94.9	81.2 95.6		

Simulation results for the rectangular confidence region for a point on the ROC curve, and simultaneous and pointwise confidence bands for the ROC curve. Note: Results are based on 10,000 runs with 60% of subjects having replicate (4) observations. Naive estimates are based on 1 randomly selected observation per subject in each group. The coverage (nominal 95% level) for the rectangular confidence region is based on the -1 transformed confidence intervals described in Section 4. The true biomarker is normal (6.5, 0.3) in diseased subjects and normal (5.7, 0.2) in non-diseased subjects. The measurement error is normal (0, τ_D^2) in each group D. The true sensitivity $Se = 0.85$, specificity $Sp = 0.7$, and disease prevalence $= 0.5$. $(\cdot)$ denotes the cumulative distribution function of a normal (0, 1) random variable.

Table 2

Type of Confidence Interval/Band (True Value)	Sample Size $n_1 = n_0$	Percent Coverage			
		Naive Method		Proposed Method	
		$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$	$\tau_D^2=0.5$
Rectangular confidence region for a point on the ROC curve (0.3, 0.85)	50	87.3	25.9	95.3	96.1
	100	79.0	4.0	95.3	96.6
	150	70.7	0.4	95.0	96.1
Pointwise confidence band for ROC(t) at $t=0.3$ (0.85)	50	83.8	16.7	95.0	96.8
	100	71.9	1.6	95.5	96.0
	150	59.9	0.1	94.8	95.8
Simultaneous confidence band for ROC(θ) (ROC($\theta = \{1.465 + 0.816 \cdot -1(\theta)\}$), t [0,1])	50	88.1	21.5	94.0	97.2
	100	77.2	1.8	95.4	95.8
	150	64.0	0.1	95.0	95.8

Table 3

Simulation Results for Positive and Negative Predictive Value. Note: Results are based on 10,000 runs with 35%, 60%, or 100% of subjects having replicate (4) observations. The coverage (nominal 95% level) is based on the logit-transformed confidence intervals described in Web Appendix C. Naive estimates are based on M observations per subject in each group. The true biomarker is normal (6.5, 0.3) in diseased subjects and normal (5.7, 0.2) in non-diseased subjects. The measurement error is normal (0, τ_D^2) in each group D. The true sensitivity Se = 0.85, specificity Sp = 0.7, and disease prevalence = 0.5. The true positive likelihood ratio (PLR) = 2.83 and negative likelihood ratio (NLR) = 0.21. The sample size in each group is 150.

Parameter (True Value)	Percent Replicates	M	Method	Bias			Standard Deviation			Mean($\widehat{se}$)			Percent Coverage		
				$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$	$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$	$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$	$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.5$
PLR (2.83)	35	1	Naive Proposed	-0.372 0.060	-0.927 0.218	0.251 0.385	0.183 0.994	0.284 0.378	0.189 0.720	74.4 95.5	3.1 96.1				
	60	1	Naive Proposed	-0.368 0.061	-0.931 0.129	0.257 0.381	0.177 0.672	0.284 0.365	0.188 0.556	74.0 95.1	2.8 96.0				
	100	4	Naive Proposed	-0.098 0.049	-0.437 0.090	0.292 0.341	0.244 0.462	0.342 0.337	0.271 0.441	95.6 95.2	64.2 95.7				
NLR (0.21)	35	1	Naive Proposed	0.064 -0.001	0.217 -0.005	0.040 0.040	0.053 0.059	0.041 0.040	0.053 0.058	57.2 95.2	0.1 96.6				
	60	1	Naive Proposed	0.063 -0.001	0.218 -0.002	0.040 0.039	0.052 0.052	0.041 0.038	0.053 0.052	58.1 94.9	0.1 96.0				
	100	4	Naive Proposed	0.018 -0.001	0.077 -0.002	0.035 0.036	0.041 0.044	0.037 0.036	0.042 0.044	92.3 95.3	44.5 95.7				

Table 4

Simulation Results for Sensitivity and Specificity. Note: Results are based on 10,000 runs with 35%, 60%, or 100% of subjects having replicate (4) observations. The coverage (nominal 95% level) is based on the τ_D^2 transformed confidence intervals described in Section 4 Naive estimates are based on M observations per subject in each group. The true biomarker is normal (6.5, 0.3) in diseased subjects and normal (5.7, 0.2) in non-diseased subjects. The measurement error is normal (0, τ_D^2) in each group D. The true sensitivity Se = 0.85, specificity Sp = 0.7, and disease prevalence p = 0.5. The true positive predictive value (PPV) = 0.74 and negative predictive value (NPV) = 0.82. The sample size in each group is 150.

Parameter (True Value)	Percent Replicates	M	Method	Bias			Standard Deviation			Mean (SE)			Percent Coverage		
				$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$	$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$	$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$	$\tau_D^2=0.1$	$\tau_D^2=0.5$	$\tau_D^2=0.1$
PPV (0.74)	35	1	Naive Proposed	-0.030 0.002	-0.085 0.006	0.021 0.024	0.021 0.024	0.021 0.039	0.023 0.025	0.023 0.024	0.022 0.038	0.022 0.038	74.4 95.5	3.1 96.1	
	60	1	Naive Proposed	-0.029 0.002	-0.085 0.003	0.021 0.024	0.021 0.024	0.021 0.033	0.023 0.024	0.023 0.024	0.022 0.033	0.022 0.033	74.0 95.1	2.8 96.0	
	100	4	Naive Proposed	-0.008 0.001	-0.035 0.003	0.021 0.022	0.021 0.022	0.021 0.028	0.024 0.022	0.024 0.022	0.023 0.028	0.023 0.028	95.6 95.2	64.2 95.7	
NPV (0.82)	35	1	Naive Proposed	-0.041 0.001	-0.124 0.006	0.024 0.027	0.024 0.027	0.026 0.041	0.025 0.027	0.025 0.027	0.026 0.040	0.026 0.040	57.2 95.2	0.1 96.6	
	60	1	Naive Proposed	-0.040 0.002	-0.125 0.003	0.025 0.026	0.025 0.026	0.025 0.035	0.025 0.026	0.025 0.026	0.026 0.035	0.026 0.035	58.1 94.9	0.1 96.0	
	100	4	Naive Proposed	-0.011 0.001	-0.048 0.002	0.023 0.024	0.023 0.024	0.025 0.030	0.024 0.024	0.024 0.024	0.025 0.030	0.025 0.030	92.3 95.3	44.5 95.7	