Published in final edited form as:
Stat Med. 2009 February 28; 28(5): 780–797. doi:10.1002/sim.3514.

Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard‡

Paul S. Albert[*,†]
Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD 20892, U.S.A

Summary

The goal in diagnostic medicine is often to estimate the diagnostic accuracy of multiple experimental tests relative to a gold standard reference. When a gold standard reference is not available, investigators commonly use an imperfect reference standard. This paper proposes methodology for estimating the diagnostic accuracy of multiple binary tests with an imperfect reference standard when information about the diagnostic accuracy of the imperfect test is available from external data sources. We propose alternative joint models for characterizing the dependence between the experimental tests and discuss the use of these models for estimating individual-test sensitivity and specificity as well as prevalence and multivariate post-test probabilities (predictive values). We show using analytical and simulation techniques that, as long as the sensitivity and specificity of the imperfect test are high, inferences on diagnostic accuracy are robust to misspecification of the joint model. The methodology is demonstrated with a study examining the diagnostic accuracy of various HIV-antibody tests for HIV.

Keywords

diagnostic error; imperfect tests; latent class models; misclassification; predictive values; prevalence; sensitivity; specificity; diagnostic accuracy

1. Introduction

In many applications in diagnostic testing, interest focuses on estimating the diagnostic accuracy of multiple binary tests relative to a reference gold standard [1]. Unfortunately, a true gold standard is often not available, leading investigators to use an imperfect reference standard (also known as an imperfect test). Various authors have discussed the bias in estimating diagnostic accuracy using an imperfect test [2–6]. Methodology has been proposed for estimating diagnostic accuracy of a single experimental test using an imperfect test when the sensitivity and specificity of the imperfect test are known [7,8] and when they are estimated from another study [9]. Two methods have been proposed for estimating diagnostic accuracy when an imperfect test does not exist or is unavailable or when the sensitivity and specificity of the imperfect test are not known or cannot be estimated. Discrepant analysis is a method in which, initially, the results of an experimental test are compared with those of an imperfect reference standard. Disagreements between these two tests are then resolved using either a gold standard test or another reference standard test. Although commonly conducted, this approach

*Correspondence to: Paul S. Albert, Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD 20892, U.S.A.
†albertp@ctep.nci.nih.gov

has been shown to be highly biased [10–15]. For multiple tests, latent class models have been proposed for estimating diagnostic accuracy [16–18], among others. These approaches involve treating the gold standard as an unobserved latent class and obtaining a model-based estimate of diagnostic accuracy relative to the unobserved gold standard test using the proposed latent class model. Latent class models have also been criticized on a number of grounds, including being sensitive to unverifiable modeling assumptions [19–21]. Specifically, Albert and Dodd [20] showed that inferences about the diagnostic accuracy of binary tests may be highly sensitive to the conditional dependence between tests (i.e. dependence between tests given the gold standard test result), yet in many practical situations, it may also be very difficult to distinguish between competing models for the dependence structure. Thus, it is important to develop alternatives to discrepancy analysis and latent class analysis when a gold standard test is not available.

Frequently, we have multiple experimental tests, an imperfect test, and estimates of sensitivity and specificity of the imperfect test relative to a gold standard test from another study. For example, in one study, investigators were interested in the diagnostic accuracy of various experimental tests for HIV. A Western blot assay is the gold standard test of true HIV status. However, this test is expensive and may not be measured in other studies. Instead, we show how to use an enzyme-linked immunoSorbent assay (ELISA) as an imperfect reference test and how to use estimates of the diagnostic accuracy of the ELISA relative to the Western blot assay (obtained from prior studies) to make valid inference on the diagnostic accuracy of the experimental tests.

This paper proposes methodology for estimating diagnostic accuracy of multiple binary tests using an imperfect gold standard test when previous estimates of sensitivity and specificity of the imperfect test relative to the gold standard test are available. This work extends the work of Baker [9], who discussed estimation of the diagnostic accuracy of a single binary test using an imperfect reference standard. Section 2 presents the methodology for estimating diagnostic accuracy of multiple tests using an imperfect reference test. In Section 3, we analyze a data set in which interest focuses on estimating diagnostic accuracy of several tests for HIV. We examine the asymptotic bias of the diagnostic accuracy estimation under misspecified dependence structures in Section 4. In Section 5, we present the results of simulation studies examining the finite sample properties of the approach. A discussion follows in Section 6.

2. Modeling Approach

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$ be J dichotomous experimental test results on subject i ($i = 1, 2, \ldots, I$), where Y_{ij} denotes the result for the jth test on subject i. We denote d_i as the results of the unobserved gold standard reference test and T_i as the imperfect test for subject i. The contribution of each individual to the likelihood of the observed test results $\mathbf{Y}_i$ and T_i is

$$P(\mathbf{Y}_i, T_i) = \sum_{l=0}^{1} P(\mathbf{Y}_i \mid T_i, d_i = l) P(T_i \mid d_i = l) P(d_i = l) \tag{1}$$

where $P(\mathbf{Y}_i|T_i, d_i)$ is the conditional distribution of the experimental tests given the imperfect reference standard T_i and the gold standard test d_i, $P(d_i = 1)$, which is denoted as P_d, is the prevalence of the gold standard test, and $P(T_i|d_i)$ characterizes the diagnostic accuracy of the imperfect relative to the gold standard test. In this approach, the results of previous studies are used to obtain estimates of $P(T_i|d_i)$, which can then be used to estimate the diagnostic accuracy of the experimental tests relative to the gold standard test. Specifically, the sensitivity and specificity of the imperfect test relative to a gold standard test, defined as $\text{SENS}_T = P(T_i = 1|$

$d_i = 1$) and $\text{SPEC}_T = P(T_i = 0|d_i = 0)$, respectively, may be estimated from a prior study of sample size N (we denote these estimated values as $\widehat{\text{SENS}}_T$ and $\widehat{\text{SPEC}}_T$, respectively).

For any conditional distribution of $Y_{ij}|T_i, d_i$, the sensitivity and specificity of the jth test can be expressed as

$$\begin{aligned} P(Y_{ij}=1|d_i=1)= \quad & P(Y_{ij}=1|T_i=1, d_i=1)\text{SENS}_T \\ & +P(Y_{ij}=1|T_i=0, d_i=1)(1-\text{SENS}_T) \end{aligned} \tag{2}$$

and

$$\begin{aligned} P(Y_{ij}=0|d_i=0)= \quad & P(Y_{ij}=0|T_i=0, d_i=0)\text{SPEC}_T \\ & +P(Y_{ij}=0|T_i=1, d_i=0)(1-\text{SPEC}_T) \end{aligned} \tag{3}$$

respectively. There are different ways to characterize the conditional distribution of $\mathbf{Y}_i|(T_i, d_i)$. The conditional independence model (IND) assumes that the multiple tests on the same individuals are independent given T_i and d_i. Specifically, the model assumes that $P(\mathrm{Y}_i|T_i, d_i)=\prod_{j=1}^{J} P(Y_{ij}|T_i, d_i)$. Under the IND model, the sensitivity and specificity of the jth test are given by (2) and (3), respectively. The Gaussian random effects (GRE) model [16] incorporates dependence across tests by assuming that $(Y_{ij}|T_i, d_i, b_i)$ is independent Bernoulli with proportion $\Phi(\beta_{jT_id_i} + \sigma_{T_id_i}b_i)$, where the random variables b_i are standard normal and Φ is the standard normal cumulative distribution function. Under the GRE model, the sensitivity and specificity of the jth test are given by (2) and (3) with $P(Y_{ij}=1|T_i, d_i=1)=E_{b_i}\left\{\Phi(\beta_{jT_i1}+\sigma_{T_i1b_i})\right\}=\Phi(\beta_{jT_i1}/\sqrt{1+\sigma^2_{T_i1}})$ and $P(Y_{ij}=0|T_i, d_i=0)=E_{b_i}\left\{1-\Phi(\beta_{jT_i0}+\sigma_{T_i0}b_i)\right\}=\Phi(-\beta_{jT_i0}/\sqrt{1+\sigma^2_{T_i0}})$, respectively [16]. Another continuous mixture model that incorporates a dependence structure similar to that of the GRE model can be formulated when the sensitivity and specificities for the J experimental tests are assumed constant (e.g. a single assay performed repeatedly). Specifically, the beta-binomial (BB) model incorporates conditional dependence by allowing individual response proportions to follow beta distributions. We assume that $(Y_{ij}|T_i, d_i, p_{T_id_i})$ are independent Bernoulli with response proportion $p_{T_id_i}$, where p_{T_i0} has a beta distribution with parameters α_{T_i0} and β_{T_i0} and where p_{T_i1} has a beta distribution with parameters α_{T_i1} and β_{T_i1}. Thus, $(\sum_{j=1}^{J} Y_{ij}|T_i, d_i=0)$ is BB with parameters α_{T0} and β_{T0} and $(\sum_{j=1}^{J} Y_{ij}|T_i, d_i=1)$ are BB with parameters α_{T_i1} and β_{T_i1}. For the BB model, the common sensitivity and specificity across the J tests are given by $\{\alpha_{11}/(\alpha_{11} + \beta_{11})\}\text{SENS}_T + \{\alpha_{01}/(\alpha_{01} + \beta_{01})\}(1 - \text{SENS}_T)$ and $\text{SPEC}_T/(\alpha_{00} + \beta_{00}) + (1 - \text{SPEC}_T)/(\alpha_{10} + \beta_{10})$, respectively.

A substantially different way to incorporate dependence between tests is with a finite mixture (FM) rather than a continuous mixture model. Specifically, Albert *et al*. [18] and Albert and Dodd [20] proposed a model in which some individuals who are truly positive are always classified as positive by any test while others are subject to diagnostic error. Similarly, some truly negative subjects are always classified as negative by any test while others are subject to diagnostic error. We propose a similar model for this problem. Let l_{T_i, d_i} be an indicator of whether the ith subject, given imperfect test T_i and disease status d_i, is always classified correctly, so that $l_{T_i1} = 1$ when a true positive subject is always rated positive and $l_{T_i0} = 1$ when

a true negative is always rated negative. Further, define $\eta_{T0} = P(l_{T_i0} = 1)$ and $\eta_{T1} = P(l_{T_i1} = 1)$. Test results Y_{ij} given T_i, d_i and $l_{T_i d_i}$ are independent Bernoulli with probability

$$P\left(Y_{ij}=1\middle|T_i, d_i, l_{T_i d_i}\right) = \begin{cases} 1 & \text{if } d_i=1 \text{ and } l_{T_i 1}=1 \\ 0 & \text{if } d_i=0 \text{ and } l_{T_i 0}=1 \\ \omega_j(T_i, 1) & \text{if } d_i=1 \text{ and } l_{T_i 1}=0 \\ 1-\omega_j(T_i, 0) & \text{if } d_i=0 \text{ and } l_{T_i 0}=0 \end{cases} \quad (4)$$

where $\omega_j(T_i, d_i)$ is the probability of the jth test making a correct diagnosis when the individual is subject to diagnostic error ($l_{T_i1} = 0$ or $l_{T_i0} = 0$) and has an imperfect test result T_i. Under the FM model, the sensitivity and specificity of the jth test are given by (2) and (3), respectively, with $P(Y_{ij} = 1|T_i, d_i = 1) = \eta_{T_i1} + (1 - \eta_{T_i1})\omega_j(T_i, 1)$ and $P(Y_{ij} = 0|T_i, d_i = 0) = \eta_{T_i0} + (1 - \eta_{T_i0})\omega_j(T_i, 0)$.

In some situations, investigators may be interested in estimating a common sensitivity and specificity across J tests (e.g. multiply repeated tests of the same type). The common sensitivity and specificity can be estimated with the previously described models under the constraint that the sensitivity and specificity are the same across tests. For example, we assume that $\beta_{1T_i d_i} = \beta_{2T_i d_i} = \cdots = \beta_{JT_i d_i}$ for the GRE model and $\omega_1(T_i, d_i) = \omega_2(T_i, d_i) = \cdots = \omega_J(T_i, d_i)$ for the FM model when estimating a common sensitivity and specificity across the J tests.

The proposed approach incorporates conditional dependence between the experimental tests and the imperfect test (i.e. $\mathbf{Y}_i$ and T_i are dependent conditional on d_i) since $P(\mathbf{Y}_i|T_i, d_i)$ depends on T. A special case of this approach is when the experimental tests and the imperfect test are conditionally independent given d_i. In this case, $P(\mathbf{Y}_i|T_i, d_i) = P(\mathbf{Y}_i|d_i)$ in (1) and the parameters in each of the conditional dependence models will not depend on T_i (i.e. suppress the subscript T_i for all parameters in each of the conditional dependence models). Under this assumption, the sensitivity and specificity of the experimental tests are simply functions of the parameters of the conditional distribution of $Y_{ij}|d_i$ and will not explicitly depend on SENS_T and SPEC_T. For example, the sensitivity and specificity under the GRE model are simply $\Phi\left(\beta_{j1}/\sqrt{1+\sigma_1^2}\right)$ and $\Phi\left(-\beta_{j0}/\sqrt{1+\sigma_0^2}\right)$, respectively. Likewise, the sensitivity and specificity under the FM model are $\eta_1 + (1 - \eta_1)\omega_j(1)$ and $\eta_0 + (1 - \eta_0)\omega_j(0)$, respectively.

There are other important special cases of this approach. When $\text{SENS}_T = \text{SPEC}_T = 1$, the approach reduces to the case in which the gold standard test is observed. When $\text{SENS}_T = \text{SPEC}_T = 0.50$, there is no information about the gold standard test obtained from observing the imperfect test, and the approach reduces to a latent class models without a gold standard. For example, when $P(\mathbf{Y}_i|T_i, d_i) = P(\mathbf{Y}_i|d_i)$ and $\text{SENS}_T = \text{SPEC}_T = 0.50$, the likelihood (1) is proportional to the standard latent class model likelihood for estimating diagnostic accuracy without a gold standard [18]. Thus, as with latent class models with no gold standard, the likelihood (1) is invariant to a relabeling of the latent variable (i.e. $d_i = 0$ may correspond to a positive as well as a negative test), and there are two solutions that maximize the likelihood. Further, as for modeling diagnostic accuracy without a gold standard, when $\text{SENS}_T = \text{SPEC}_T = 0.50$, $J \geqslant 5$ is required for identifiability in order to estimate a common sensitivity and specificity across J tests with a GRE, FM, or BB model, and $J \geqslant 4$ is required for identifiability in order to estimate test-specific diagnostic accuracy using a GRE or FM model [20].

When estimating the sensitivity and specificity of each test (i.e. test-specific estimators of diagnostic accuracy), an alternative to jointly modeling the multiple tests is to formulate a

separate model for each test. This is done by modeling the joint distribution of Y_{ij} and T_i for each test, where an individual's contribution to the likelihood is

$$P(Y_{ij}, T_i) = \sum_{l=0}^{1} P(Y_{ij} | T_i, d_i = l) P(T_i | d_i = l) P(d_i = l) \tag{5}$$

where $P(Y_{ij}|T_i, d_i)$ along with expressions (2) and (3) characterizes the sensitivity and specificity for the jth test. As in the joint modeling of $\mathbf{Y}_i$ and T_i, $P(T_i|d_i)$ is estimated from a previous study ($\widehat{\text{SENS}}_T$ and $\widehat{\text{SPEC}}_T$). Further, as in the joint modeling, this approach simplifies when $P(\mathbf{Y}_i|T_i, d_i) = P(\mathbf{Y}_i|d_i)$. The sensitivity and specificity for the jth test can be estimated by maximizing (5) separately for each test. Although this approach does not make assumptions about the dependence between experimental tests, it is likely highly inefficient relative to a joint modeling approach. We will explore this in more detail in Sections 3 and 5.

Various authors have criticized the use of sensitivity and specificity as a measure of diagnostic accuracy [22-24] and references within. Much of this criticism focuses on the facts that sensitivity and specificity have limited clinical utility and that, in most practical situations, a single diagnostic test will be insufficient for adequately assessing diagnostic accuracy. These and other authors have suggested the use of post-test probabilities or predictive values, which can be easily generalized to incorporate multivariate factors in assessing diagnostic accuracy. We can estimate the multivariate predictive value of the multiple experimental tests $\mathbf{Y}_i$ using (1), by noting that

$$P(d_i = 1 | \mathbf{Y}_i) = \frac{\sum_{t=0}^{1} P(\mathbf{Y}_i | T_i = t, d_i = 1) P(T_i | d_i = 1) P(d_i = 1)}{\sum_{d=0}^{1} \sum_{t=1}^{1} P(\mathbf{Y}_i | T_i = t, d_i = d) P(T_i | d_i = d) P(d_i = d)} \tag{6}$$

For parameter estimation, the likelihood $\log L = \sum_{i=1}^{I} \log L_i$ is maximized, where L_i is given by (1) or (5), and where $P(T_i|d_i)$ is based on an external data source (often based on an estimate provided in a prior study). The likelihood was maximized using a quasi-Newton Raphson algorithm in GAUSS [25]. Although asymptotic standard errors for the parameters and model-based estimates of sensitivity and specificity can be derived using delta method approximations (Baker discussed this for the case of a single experimental test [9]), we estimated standard errors using the bootstrap [26] where an individual's data were resampled with replacement and the standard deviations of the resulting bootstrap estimates were estimated (800 bootstrap realizations). The uncertainty due to the estimates of $P(T_i|d_i)$ from the external prior study (with sample size N) can easily be incorporated by generating, for each bootstrap realization, $P\hat{}(T_i = l|d_i = l) = W_l/N_l$, where N_1 is the number of gold standard test results with $d_i = 1$ in the prior study which is generated as bin(N, P_d) where P_d is estimated from the prior study, N_0 is the number of gold standard test results with $d_i = 0$ which is calculated as $N_0 = N - N_1$ and $W_l \sim \text{bin}\left(N_l, \widehat{\text{SENS}}_T^{\,l}\, \widehat{\text{SPEC}}_T^{\,l-l}\right)$. Software for fitting these models is written in GAUSS and is available from the author on request.

3. Application

We apply this methodology to estimate the diagnostic accuracy of three assays for detecting HIV. Specifically, we examined the diagnostic accuracy of ag121, p24, and gp120 using an ELISA as an imperfect gold standard with a data set of 428 samples provided by Alvord *et*

al. [27]. Prior to the publication of this study, another study estimated the sensitivity and specificity of the ELISA as being 97.7 per cent (86 of 88) and 92.6 per cent (275 of 297) for detecting HIV [28]. Using this information, we estimated the diagnostic accuracy of ag121, p24, and gp120 using the ELISA as the imperfect reference standard. Initially, we fit the IND, GRE, and FM models assuming that $P(\mathbf{Y}_i|T_i, d_i) = P(\mathbf{Y}_i|d_i)$. Table I presents the results of a model using the imperfect test (Equation (1)) and an IND, GRE, and FM dependence structure to the data set of Alvord *et al.* Estimates of sensitivity and specificity and bootstrap standard errors are almost identical across the IND, GRE, and FM conditional dependence structures. This suggests that estimation (estimates and variances) may be robust to the choice of dependence between tests. Further, the log-likelihoods were −634.19, −627.21, and −627.21 for the IND, GRE, and FM models, respectively, suggesting that (i) the two models that account for conditional dependence fit better than the model that assumes conditional independence and (ii) it is difficult to distinguish between the GRE and FM models. (We cannot formally compare the GRE and FM models using a likelihood ratio test since neither model is nested within the other.) We also fit joint models which do not assume that Y_{ij} and T_i are conditionally independent given d_i ($P(\mathbf{Y}_i|T_i, d_i) \neq P(\mathbf{Y}_i|d_i)$). Likelihood ratio tests comparing the simpler model with the more complex model, which allows the conditional distribution to depend on T_i, were not significant for either the FM or GRE models. Further, the estimated sensitivities and specificities of the three tests for this more complex model were similar to those presented in Table I (data not shown).

We also estimated the test-specific sensitivity and specificity by modeling each experimental test separately. This was done by maximizing (5) separately for the ag121, p24, and gp120 assays (Table I) and assuming $P(\mathbf{Y}_i|T_i, d_i) = P(\mathbf{Y}_i|d_i)$. These results are close to approaches where the experimental tests are jointly modeled (IND, GRE, and FM dependence structures) with slightly increased standard errors. As with the joint model, estimates obtained with the more complex model that allowed $P(\mathbf{Y}_i|T_i, d_i)$ to depend on T_i resulted in estimates similar to those presented in Table I.

Alvord's data contain the actual HIV status for all 428 patients, which can be used to examine the performance of the approach. Using the gold standard HIV status, the sensitivity and specificity were, respectively, 0.99 (*SE* = 0.006) and 0.98 (0.01) for the ag121 test, 0.56 (0.03) and 0.97 (0.01) for p24, 0.89 (0.02) and 1.00 (0) for gp120, and 0.98 (0.01) and 0.93 (0.02) for an ELISA test. These results are all similar to those estimates obtained by using the imperfect reference test with the proposed methodology. Further, the sensitivity and specificity estimates for the ELISA in this study are similar to those estimated from the previous study [28].

We estimated the multivariate predictive values for all combinations of the three experimental tests by using (6). These were estimated under the GRE and FM models and compared with the predictive values using the actual HIV status described in the previous paragraph. We report predictive values for combinations of tests with more than five patients. When all three tests were negative, the predictive value was 0 for both models. This is close to the predictive value using the actual HIV status, which was $\frac{2}{183}=0.01$. When all three tests were positive, the predictive value was 1 for both models as compared with $\frac{128}{128}=1.0$ using the actual HIV status. When p24 was negative and the remaining two tests were positive, the predictive value under both models was 1.0 as compared with $\frac{83}{83}=1.0$ using the actual HIV status. Finally, when ag121 was negative and the remaining two tests were positive, the predictive value for both models was 0.82 as compared with $\frac{18}{21}=0.86$ using the actual HIV status.

The data analysis poses an intriguing question, namely, are inferences made with the joint modeling approach which incorporates the diagnostic accuracy of the imperfect reference test robust to misspecification of the dependence between tests? Albert and Dodd [20] showed that when estimating diagnostic accuracy without a gold standard (which is a special case of the model when $SENS_T = SPEC7_T = 0.5$), it is difficult to distinguish between competing models for the dependence between tests and the choice of dependence structure has a large impact on sensitivity and specificity estimation (i.e. lack of robustness to the dependence structure between tests). Is the problem alleviated when we incorporate information about the diagnostic accuracy of the imperfect test from a previous study into the analysis? We examine this with analytical and simulation techniques in the following two sections.

4. Asymptotic Results

We examined the asymptotic bias ($I \to \infty$) in estimating diagnostic accuracy and prevalence using an imperfect test when the dependence structure is misspecified. These asymptotic biases are also calculated under the assumption that the sensitivity and specificity of the imperfect test relative to the gold standard test are estimated from a large sample (i.e. $N \to \infty$ with $\widehat{SENS}_T = SENS_T$ and $\widehat{SPEC}_T = SPEC_T$). Initially, we considered the bias for the case when a common sensitivity and specificity are estimated across J tests (this is most appropriate when the same assay is repeated multiple times). The misspecified maximum-likelihood estimator for the model parameters, denoted by $\hat{\boldsymbol{\theta}}^*$, converges to the value $\boldsymbol{\theta}^*$, where

$$\theta^* = \arg\max_{\theta} E_{\mathrm{Tr}} \left[\log_M L(\mathrm{Y}_i, T_i, \theta) \right] \tag{7}$$

and $\log L_M(\mathbf{Y}_i, T_i, \boldsymbol{\theta})$ is the individual contribution to the log-likelihood under the assumed model M and the expectation is taken under the true model Tr. The notation

$$E_{\mathrm{Tr}}(\log L_M) = E_{\mathrm{Tr}} [\log L(\mathrm{Y}_i, T_i, \theta)]\big|_{\theta=\theta^*} \tag{8}$$

denotes the expectation (taken under the true model Tr) of an individual's contribution to the log-likelihood under the assumed model M when evaluated at $\boldsymbol{\theta}^*$. Sensitivity and specificity are model-dependent functional forms of the model parameters, $SENS^* = g_1(\boldsymbol{\theta}^*)$ and $SPEC^* = g_2(\boldsymbol{\theta}^*)$, where g_1 and g_2 relate model parameters to sensitivity and specificity. Estimators of sensitivity, specificity, and prevalence converge to SENS*, SPEC*, and P_d^*, respectively, under misspecified models. Expressions for an individual's contribution to the expected log-likelihood under the correct and misspecified models are provided in Appendix A. Asymptotic biases for sensitivity, specificity, and prevalence are defined as SENS*−SENS, SPEC*−SPEC, and $P_d^* - P_d$, respectively.

Based on results described by Tan *et al.* [29] and Heagerty and Kurland [30], who showed that marginal quantities in generalized linear mixed models (such as sensitivity, specificity, and prevalence) are nearly unbiased when the random effects distribution is misspecified, we would expect that estimators of sensitivity, specificity, and prevalence would be nearly unbiased under a misspecified model when $SENS_T = SPEC_T = 1$. Further, based on results described by Albert and Dodd [20], who showed that estimation of diagnostic accuracy and prevalence using latent class models without a gold standard are highly sensitive to the dependence structure between tests, we would expect that when $SENS_T = SPEC_T = 0.50$, estimators of diagnostic accuracy and prevalence would be asymptotically biased under a misspecified model. Table II shows

the asymptotic bias for estimating prevalence and the common diagnostic accuracy across $J = 5$ tests when a GRE model is assumed when the true model is an FM model. The asymptotic calculations are for the case in which the sensitivity, specificity, and prevalence are 0.75, 0.90, and 0.20, respectively, and $P(\mathbf{Y}_i = 1|T_i, d_i) = P(\mathbf{Y}_i|d_i)$, but the results apply more broadly. As expected, estimators under the misspecified GRE model were unbiased when $SENS_T = SPEC_T = 1$ and highly biased when $SENS_T = SPEC_T = 0.50$. Table II shows the biases for values of $SENS_T$ and $SPEC_T$ between 0.50 and 1. The asymptotic biases under the misspecified GRE model were small for values of $SENS_T$ and $SPEC_T$ above 0.75. Thus, in this case, estimation is robust to model misspecification under practical values of the sensitivity and specificity of the imperfect reference test. Table III also presents individual contributions to the expected values of the log-likelihoods under the misspecified and correct models for the dependence between tests. When $SENS_T = SPEC_T = 0.50$, the expected log-likelihood for the misspecified GRE model is almost the same as the expected log-likelihood for the correctly specified model. Thus, in this case, it may be very difficult to distinguish between competing models for the dependence between tests. Table II shows that as $SENS_T$ or $SPEC_T$ increases over 0.50, the expected log-likelihood for the correctly specified model becomes larger than the expected log-likelihood under the misspecified model, suggesting that it will become easier to distinguish between competing models. This will be examined in more detail in the simulation results presented in Section 5.

There are cases when the asymptotic biases are not negligible for reasonably high values of $SENS_T$ and $SPEC_T$. In these cases, the expected individual contributions to the log-likelihood under the correctly specified models were considerably larger than the expected contributions under the misspecified models, making it possible to distinguish between models. For example, when the true model is a GRE model ($\sigma_0 = \sigma_1 = 1$), SENS = 0.75, SPEC = 0.90, $P_d = 0.20$, and $J = 5$, estimates of sensitivity, specificity, and prevalence under an FM model converge to SENS* = 0.81, SPEC* = 0.91, and $P_d^* = 0.19$ when $SENS_T = SPEC_T = 0.90$. The expected individual contribution to the log-likelihood under the correct GRE model was $E_{GRE}[\log L_{GRE}] = -2.3897$ as compared with the expected log-likelihood under the misspecified FM model of $E_{GRE}[\log L_{FM}] = -2.4000$.

The FM and the GRE models are very different ways to incorporate dependence between tests. Of interest is examining the effect of model misspecification on estimation when the models are much more similar in how they incorporate dependence between tests (in this case it may be very difficult to distinguish between competing models for the dependence structure even when the sensitivity and specificity of the imperfect test are 1). The BB and the GRE models are very similar in that each uses a continuous mixture distribution to incorporate dependence. Figure 1 shows contour plots for the relative asymptotic bias of sensitivity, specificity, and prevalence of an experimental test ($J = 5$ repeated tests) for different known sensitivities and specificities of the imperfect relative to the gold standard test. As we expected from results on latent class models without a gold standard [20], there is sizable bias when $SENS_T$ and $SPEC_T$ are close to 0.50 (corresponding to the no-gold-standard case). However, the asymptotic bias is negligible for even a slight increase over 0.50 in either the sensitivity or specificity of the imperfect test.

Table III shows the asymptotic bias of test-specific sensitivity and specificity for four tests under a misspecified dependence structure with various sensitivities and specificities of the imperfect reference test relative to the gold standard test. Specifically, we assume a GRE model when the true model is an FM model. For simplicity, we assume that $P(\mathbf{Y}_i|T_i, d_i) = P(\mathbf{Y}_i|d_i)$, but the results will not be sensitive to this fact. As with the common sensitivity and specificity calculations, estimates are unbiased when $SENS_T = SPEC_T = 1$ and estimates are highly biased when $SENS_T = SPEC_T = 0.50$. Test-specific sensitivity and specificity have little bias when the sensitivity and specificity of the imperfect reference test are each above 0.75. Further,

although the individual contributions of the expected log-likelihoods under a misspecified and correct model are nearly identical when $SENS_T = SPEC_T = 0.50$, they begin to more substantially differ as the sensitivity and specificity increase. This suggests that it will be increasingly easier to distinguish between these competing models as the sensitivity and specificity of the imperfect test increase.

The following section shows the finite sample properties of the methodology using simulation studies.

5. Simulations

We examine the finite sample properties under the assumption of a common sensitivity and specificity across multiple tests and under the assumption of different test-specific sensitivities across tests. Table IV shows results for a common sensitivity and specificity with five tests when the true model is an FM model and when the assumed model is a GRE model for a sample size of 1000 individuals ($I = 1000$). For simplicity we assume that for both the true and assumed models $P(\mathbf{Y}_i|T_i, d_i) = P(\mathbf{Y}_i|d_i)$, but the results should apply more broadly. The results show little bias under model misspecification when the sensitivity and specificity of the imperfect test relative to the gold standard test are above 75 per cent. Further, although it is difficult to distinguish between the very different GRE and FM models when $SENS_T = SPEC_T = 0.50$ (the log-likelihood for the correct model is greater than one unit larger than the log-likelihood for the misspecified model in only 11 per cent of simulated realizations), it is much easier to distinguish between models when the sensitivity and specificity of the imperfect test are greater than 75 per cent (the log-likelihood under the correct model is greater than one unit larger than that under the incorrect model in over 40 per cent of the simulations). Thus, in this case, when the diagnostic accuracy of the imperfect relative to the gold standard test is high, estimation is robust to model misspecification and we are able to distinguish between competing models with highk probability.

There is a more sizable bias when the true model is a GRE and the misspecified model is an FM model and either the sensitivity or specificity is below one. For example, a simulation was conducted as in Table IV but with data generated with a GRE model with $\sigma_0 = \sigma_1 = 1$, SENS = 0.75, SPEC = 0.90, and $P_d = 0.20$ with $SENS_T = SPEC_T = 0.90$. Average estimates of $\widehat{SENS}$, $\widehat{SPEC}$, and $\widehat{P_d}$ were 0.80, 0.90, and 0.19, respectively, demonstrating a sizable finite sample bias for estimating sensitivity. However, the log-likelihood under the correct GRE model was larger (by at least one unit) than the log-likelihood of the incorrect FM model in greater than 99 per cent of simulated data sets.

Table V focuses on the properties of multivariate predictive values given by (6). Specifically, we estimate the probability of the true disease status being positive given the sum of the five tests. These simulations were identical to those presented in Table IV. Under the correctly specified FM model, the estimated predictive values were nearly unbiased for all values of $SENS_T$ and $SPEC_T$. However, there was bias under the misspecified GRE model. This bias was negligible when $SENS_T = SPEC_T = 1$ but increased as both $SENS_T$ and $SPEC_T$ approached 0.5, with the bias being very substantial when $SENS_T = SPEC_T = 0.50$. In addition, even when $SENS_T$ and $SPEC_T$ were both near 1, there was a substantial efficiency gain in using the correct joint model for estimating the multivariate predictive values.

The simulations assumed that $SENS_T$ and $SPEC_T$ are known values (this corresponds to using a very large study to estimate the sensitivity and specificity of the imperfect test relative to the gold standard test). Table VI shows the results of simulations under a correctly specified FM model with $SENS_T$ and $SPEC_T$ estimated from a prior study with varying sample size. As for the simulations with known values of $SENS_T$ and $SPEC_T$, the results show that estimates of

sensitivity and specificity are unbiased under a correctly specified joint model even when a relatively small study is used to estimate the diagnostic accuracy of the imperfect reference test. Further, the variability in $\widehat{\text{SENS}}$, $\widehat{\text{SPEC}}$, and $\widehat{P_d}$ (squared-standard errors in the table) does not decrease proportionally to sample size. The limiting factor on the variability of these estimates is the size of the current study and not the size of the study that estimates SENS_T and SPEC_T. Thus, there is little advantage of a large versus moderate-size prior study for estimating the diagnostic accuracy of the imperfect test.

Table VII shows the results for estimating the test-specific diagnostic accuracy for four tests when the true model is an FM model and the GRE model is the assumed model (again, we assume that $P(\mathbf{Y}_i|T_i, d_i) = P(\mathbf{Y}_i|d_i)$ in this simulation). Similar to the simulation results for a common sensitivity and specificity, the robustness under a misspecified model improves as the diagnostic accuracy of the imperfect reference test increases. When $\text{SENS}_T = \text{SPEC}_T = 1$, the results (mean estimates and standard deviations) are almost identical under the correct FM and incorrect GRE models. Estimation is robust to the dependence between tests as long as the sensitivity and specificity of the imperfect test are relatively high (at or above 75 per cent). Further, the log-likelihood is larger for the correctly specified model in 96, 91, and 88 per cent of simulated data sets when the sensitivity and specificity of the imperfect test are both 1, 0.90, and 0.75, respectively. This is in contrast to large biases under a misspecified dependence structure reported by Albert and Dodd [20] for latent class models without a gold standard (corresponding to $\text{SENS}_T = \text{SPEC}_T = 0.50$). For an identical model configuration, they showed (in their Table VI) that there are large finite sample biases for a GRE model when the true dependence structure is an FM model. Further, they demonstrated that the log-likelihood under a correctly specified FM model was larger than the log-likelihood under a misspecified GRE model in only 63 per cent of the simulated data sets.

Table VII also presents average estimates and standard errors of test-specific sensitivity and specificity obtained by fitting model (5) separately for each test. When $\text{SENS}_T = \text{SPEC}_T = 1$, estimates obtained by estimating the diagnostic accuracy individually by test are essentially the same as those obtained using the correct model. Although estimates obtained with individual tests make no assumptions about the dependence between tests, they are substantially more variable than estimates obtained with the joint modeling approach when SENS_T or SPEC_T is less than 1. For example, when $\text{SENS}_T = \text{SPEC}_T = 0.9$, estimates are on average greater than 50 per cent more efficient with the correctly specified FM model as compared with estimating diagnostic accuracy separately for each of the four tests. This efficiency gain is even larger when $\text{SENS}_T = \text{SPEC}_T = 0.75$.

6. Discussion

We proposed methodology for estimating the diagnostic accuracy of multiple binary tests using an imperfect reference test when information about the diagnostic accuracy of the imperfect relative to the gold standard test is available from a prior study. The application we considered concerns estimating experimental screening tests for HIV using an ELISA as the imperfect reference test. A prior study was used to estimate the sensitivity and specificity of the ELISA relative to a gold standard assessment of HIV status. The resulting estimates of sensitivity and specificity for the experimental tests were nearly identical to those obtained with the actual gold standard test in the current study, providing evidence that the methodology may be useful.

We demonstrated using analysis, simulations, and analytical techniques that inferences on diagnostic accuracy were robust to the dependence between experimental tests as long as the sensitivity or specificity of the imperfect test was moderately high. The robustness was found for individual test sensitivity and specificity as well as for multivariate predictive values.

However, particularly for estimating multivariate predictive values, there may be substantial efficiency advantages to correctly specifying the joint model for the dependence between tests.

The joint modeling approach is necessary when estimating a common sensitivity and specificity across *J* tests (a common situation when the same test is repeated multiple times). However, when estimating test-specific sensitivity and specificity (i.e. different sensitivity and specificity for each test), estimation can be done separately for each test using (5) as an alternative to joint modeling. Although separate estimation clearly does not make assumptions about the dependence between tests, in most situations, it is highly inefficient relative to joint modeling. Thus, there is an inherent trade-off between the gain in efficiency of joint modeling versus the potential bias if the dependence between tests is misspecified. I favor the use of the joint model since our asymptotic bias and simulation studies suggest (under extreme model misspecification) that, in most cases, there is limited bias when the dependence structure in the joint model is misspecified.

In some applications, a series of prior studies may provide estimates of diagnostic accuracy of the imperfect reference test. In this situation, a common diagnostic accuracy of the imperfect test can be estimated by weighting inversely proportional to sample size and using this common estimate with (1) or (5) to obtain estimates of the diagnostic accuracy of the experimental tests. If the diagnostic accuracies are assumed to vary across studies, then a random effects model could be employed in order to estimate a distribution of sensitivities and specificities across studies. This distribution along with (1) or (5) can then be used to estimate the distribution of likely values of the sensitivity and specificity of the experimental tests.

This paper proposes methodology for estimating the diagnostic accuracy of multiple binary tests, which incorporates information about the diagnostic accuracy of an imperfect test that may be available from external data sources. This approach was motivated by the failure of alternative approaches such as discrepancy and latent class analyses, which do not use any information about the gold standard test and have been found to be problematic. The proposed approach may be useful in early-phase diagnostic studies where measuring the gold standard test may not be feasible but when there is a sizable literature on the diagnostic accuracy of an easily obtainable imperfect reference standard. Clearly, more definitive studies with gold standard tests and clinical endpoints will be necessary to validate any promising new diagnostic tests identified with the proposed approach [31]. These more definitive studies are often subject to verification bias, since the gold standard test may be invasive and may be measurable on only a subset of highly selected patients (i.e. only those who test positive in one or more of the experimental tests). Problems of verification bias are well known [5,32], and various approaches have been proposed for estimating diagnostic accuracy when gold standard evaluation is observed on only a fraction of patients [33–36].

Appendix

Appendix A: Expected Log-Likelihood of A Misspecified Model M for A True Model Tr

Each individual's contribution to the expected log-likelihood under a true model Tr is

$$E_{\mathrm{Tr}}\left[\log L_M(\mathrm{Y}_i, T_i, \theta)\right] = \sum_{i_1=0}^{1}\sum_{i_2=0}^{1}\cdots\sum_{i_J=0}^{1}\sum_{t=0}^{1}\left[\log\left\{\sum_{l=0}^{1} P_M(\mathrm{Y}_i=(i_1, i_2, \ldots, i_J)\prime \mid T_i=t, d_i=l) \times P(T_i=t \mid d_i=l)\, P_M(d_i=l)\right\} \times \left\{\sum_{l=0}^{1} P_{\mathrm{Tr}}(\mathrm{Y}_i=(i_1, i_2, \ldots, i_J)\prime \mid T_i=t, d_i=l)\, P(T_i=t \mid d_i=l)\, P_{\mathrm{Tr}}(d_i=l)\right\}\right]$$

where $P_{\mathrm{Tr}}(\mathbf{Y}_i|T_i, d_i)$ and $P_{\mathrm{Tr}}(d_i)$ are the true distribution of $\mathbf{Y}_i$ conditional on d_i and prevalence, respectively. The misspecified conditional distribution $P_M(\mathbf{Y}_i|d_i)$ and prevalence $P_M(d_i)$ are characterized by the parameter vector $\boldsymbol{\theta}$. The diagnostic accuracy of the imperfect test T_i relative to the gold standard test d_i is characterized by $P(T_i|d_i)$, which is assumed known from a previous study.

References

1. Zhou, XH.; McClish, DK.; Obuchowski, NA. Statistical Methods in Diagnostic Accuracy. Wiley; New York: 2002.
2. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy. Annals of Internal Medicine 2004;140:189–202. [PubMed: 14757617]
3. Valenstein PN. Evaluating diagnostic tests with imperfect standards. American Journal of Clinical Pathology 1990;93:252–258. [PubMed: 2405632]
4. Boyko EJ, Alderman BW, Baron AE. Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. Journal of General Internal Medicine 1988;3:476–481. [PubMed: 3049969]
5. Begg CB. Biases in the assessment of diagnostic tests. Statistics in Medicine 1987;6:411–423. [PubMed: 3114858]
6. Qu Y, Hadgu A. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference standard. Journal of the American Statistical Association 1998;93:920–928.
7. Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. American Journal of Epidemiology 1966;88:593–602. [PubMed: 5932703]
8. Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. Journal of Chronic Diseases 1981;34:599–610. [PubMed: 6458624]
9. Baker SG. Evaluating a new test using a reference test with estimated sensitivity and specificity. Communications in Statistics: Theory and Methods 1991;20:2739–2752.
10. Hadgu A. The discrepancy in discrepant analysis. Lancet 1996;348:592–593. [PubMed: 8774575]
11. Hadgu A. Bias in the evaluation of DNA-amplification essay for detecting chlamydia trachomatis. Statistics in Medicine 1997;17:1391–1399. [PubMed: 9232760]
12. Green TA, Black CM, Johnson RE. Evaluation of bias in diagnostic-test sensitivity and specificity estimates computed by discrepant analysis. Journal of Clinical Microbiology 1998;36:375–381. [PubMed: 9466744]
13. Miller WC. Bias in discrepant analysis: when two wrongs don't make a right. Journal of Clinical Epidemiology 1998;51:219–231. [PubMed: 9495687]
14. Miller WC. Editorial response: can we do better than discrepant analysis for new diagnostic test evaluation? Clinical Infectious Diseases 1998;27:1186–1193. [PubMed: 9827267]
15. Hadgu A. Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. Journal of Clinical Epidemiology 1999;52:1231–1237. [PubMed: 10580787]
16. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics 1996;52:797–810. [PubMed: 8805757]
17. Hui SL, Zhou XH. Evaluation of diagnostic tests without a gold standard. Statistical Methods in Medical Research 1998;7:354–370. [PubMed: 9871952]
18. Albert PS, McShane LM, Shih JH, The U.S. National Cancer Institute Bladder Tumor Marker Network. Latent class modeling approaches for assessing diagnostic error without a gold standard:

with applications to p53 immunohistochemical assays in bladder tumors. Biometrics 2001;57:610–619. [PubMed: 11414591]

19. Alonzo A, Pepe M. Using a combination of reference tests to assess the accuracy of a diagnostic test. Statistics in Medicine 2001;18:2987–3003. [PubMed: 10544302]
20. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics 2004;60:427–435. [PubMed: 15180668]
21. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. Biostatistics 2006;8:474–484. [PubMed: 17085745]
22. Moons KG, Van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood-ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8:12–17. [PubMed: 9116087]
23. Moons KG, Harrell FE. Sensitivity, specificity should be de-emphasized in diagnostic accuracy studies. Academic Radiology 2003;10:670–672. [PubMed: 12809422]
24. Moons GM, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. Clinical Chemistry 2004;50:473–476. [PubMed: 14981027]
25. Aptech Systems. Gauss Systems Version 3.0 Kent. Aptech Systems; Washington: 1992.
26. Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. Chapman & Hall; New York: 1993.
27. Alvord WG, Drummond JE, Arthur LO, Biggar RJ, Goedert JJ, Levine PH, Murphy EL, Weiss SH, Blattner WA. A method for predicting individual HIV infection status in the absence of clinical information. AIDS Research and Human Retroviruses 1988;4:295–304. [PubMed: 3207513]
28. Weiss SH, Goedert JJ, Sarngadharan MG, Bodner AJ, The AIDS Seroepidemiology Collaborative working Group. Gallo RC, Blattner WA. Screening tests for HTLVI-III (AIDS Agent) antibodies: specificity, sensitivity, and applications. Journal of the American Medical Association 1985;253:221–225. [PubMed: 2981369]
29. Tan M, Qu Y, Rao JS. Robustness of the latent variable model for correlated binary data. Biometrics 1999;55:258–263. [PubMed: 11318164]
30. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalized linear mixed models. Biometrika 2001;88:973–985.
31. Biesheuvel CJ, Grobbee DE, Moons KGM. Distraction from randomization in diagnostic research. Annals of Epidemiology 2006;16:540–544. [PubMed: 16386925]
32. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. The New England Journal of Medicine 1978;299:926–930. [PubMed: 692598]
33. Begg CD, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:207–215. [PubMed: 6871349]
34. Walter SD. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. Epidemiology 1999;10:67–72. [PubMed: 9888282]
35. Albert PS, Dodd LE. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. Journal of the American Statistical Association 2008;103:61–73.
36. Bohning D, Patilea V. A capture-recapture approach for screening using two diagnostic tests with availability of disease status for the test positives only. Journal of the American Statistical Association 2008;103:212–221.

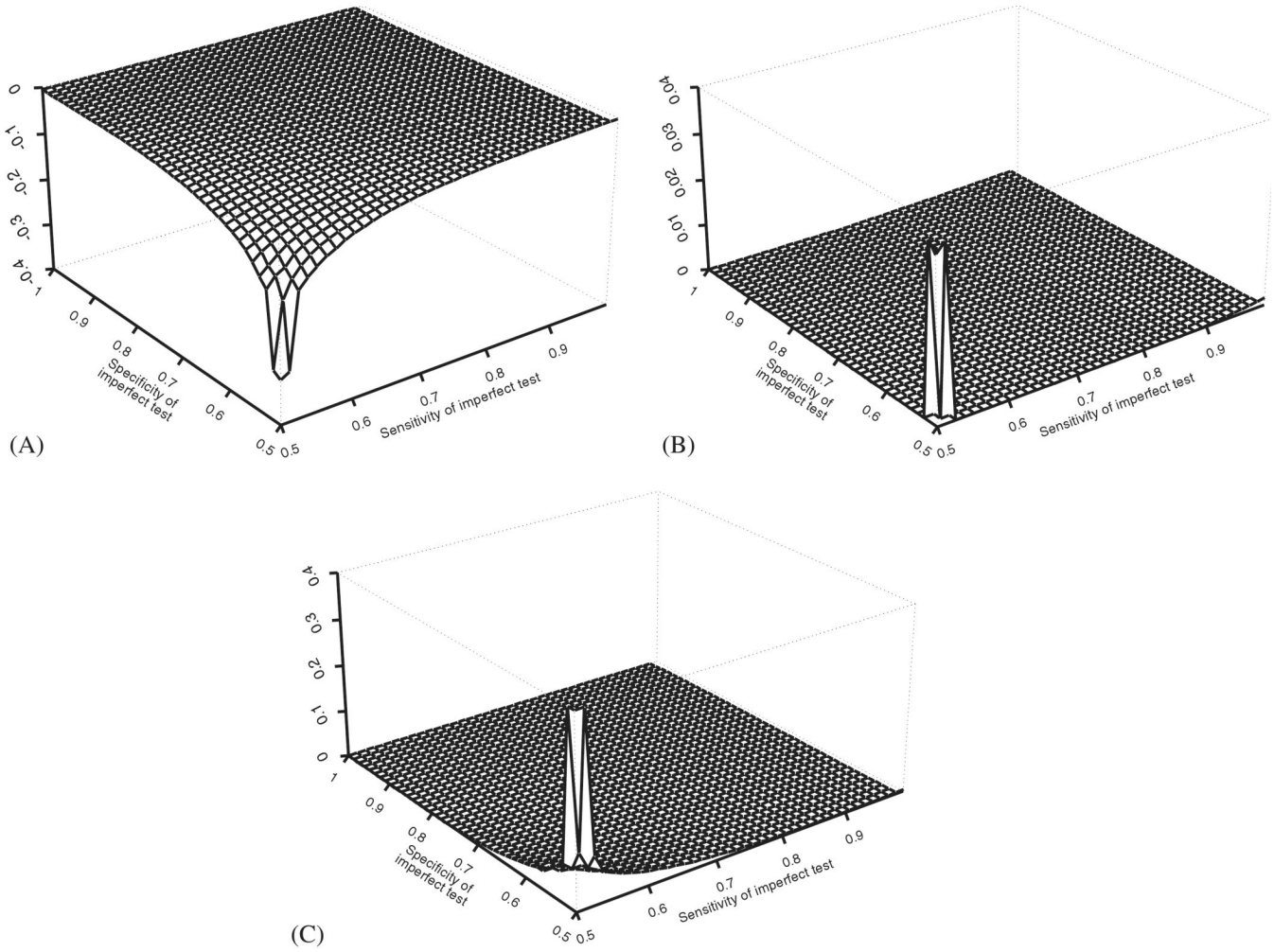

Figure 1.

The asymptotic bias of sensitivity, specificity, and prevalence of the experimental test (note that we are considering the sensitivity and specificity to be constant across the $J = 5$ tests in these calculations) when the true model is a beta-binomial (with stated model parameters) and when the incorrectly specified model (i.e. working model) is a GRE model is shown. In the notation described in Section 2, we assume that the beta-binomial model has SENS = 0.75 and SPEC = 0.90, with $\alpha_{00} = \alpha_{10} = 1$, $\beta_{00} = \beta_{10} = 0.33$, $\alpha_{01} = \alpha_{11} = 1$, $\beta_{01} = \beta_{11} = 0.11$ (i.e. (SENS* − SENS)/SENS, (SPEC* − SPEC)/SPEC, and $(P_d^* - P_d)/P_d$). Panels A, B, C show contour plots corresponding to the relative asymptotic bias for sensitivity, specificity, and prevalence, respectively. For each panel, the *x*-axis, *y*-axis, and *z*-axis reflect the sensitivity of the imperfect test, the specificity of the imperfect test, and the relative asymptotic bias of the sensitivity, specificity, and prevalence, respectively.

Table I

Estimating diagnostic accuracy and prevalence of ag121, p24, and gp120 using the imperfect ELISA.

Assay	Estimate	IND	GRE	FM	INV
ag121	SENS	1.00 (<0.001)	1.00 (0.008)	1.00 (<0.001)	1.00 (0.014)
	SPEC	0.98 (0.015)	0.98 (0.013)	0.98 (0.013)	1.00 (0.013)
p24	SENS	0.57 (0.033)	0.57 (0.034)	0.57 (0.035)	0.57 (0.035)
	SPEC	0.97 (0.014)	0.96 (0.013)	0.98 (0.014)	0.98 (0.017)
gp120	SENS	0.91 (0.020)	0.91 (0.021)	0.91 (0.020)	0.91 (0.028)
	SPEC	1.00 (<0.001)	1.00 (<0.001)	1.00 (0.001)	1.00 (<0.001)
	P_d	0.54 (0.024)	0.54 (0.025)	0.54 (0.024)	0.56* (0.025)

Diagnostic accuracy was estimated using 428 samples [27] and using prior estimates of the diagnostic accuracy of the ELISA relative to the gold standard Western blot assay [28]. Estimates were computed under conditional independence (IND), Gaussian random effects (GRE), and finite mixture (FM) models using (1). Further, we estimated tests individually (INV) using (5). Estimates and (standard errors) are presented, where standard errors were estimated using the bootstrap with 800 bootstrap samples.

*This is the prevalence estimate using only the ag121 assay. Estimates and standard errors for P_d using p24 or gp120 were nearly identical to those presented for the ag121 assay.

Table II

Asymptotic bias of sensitivity (SENS), specificity (SPEC), and prevalence (P_d) when the truth is a finite mixture (FM) model and a Gaussian random effects (GRE) is assumed.

$\mathbf{SENS}_T$	$\mathbf{SPEC}_T$	SENS*	SPEC*	P_d^*	$E_{\mathrm{FM}}[\log L_{\mathrm{GRE}}]$	$E_{\mathrm{FM}}[\log L_{\mathrm{FM}}]$
1	1	0.75	0.90	0.20	−2.3541	−2.3517
0.90	1	0.75	0.90	0.20	−2.3928	−2.3905
1	0.90	0.74	0.90	0.20	−2.5758	−2.5741
0.90	0.90	0.73	0.90	0.20	−2.6172	−2.6158
0.90	0.75	0.72	0.90	0.21	−2.7877	−2.7868
0.75	0.90	0.73	0.90	0.20	−2.6469	−2.6458
0.75	0.75	0.72	0.90	0.21	−2.8210	−2.8202
0.50	0.50	0.45	0.92	0.41	−2.9342	−2.9342

It is assumed that $J = 5$, $\eta_0 = \eta_1 = 0.20$, SENS = 0.75, SPEC = 0.90, and $P_d = 0.20$. Values of SENS*, SPEC*, and are presented for different values of the sensitivity and specificity of the imperfect reference test, SENS_T and SPEC_T. Asymptotic biases of SENS, SPEC and P_d are SENS*−SENS, SPEC*−SPEC, and $P_d^* - P_d$, respectively.

Table III

Asymptotic bias of test-specific sensitivity and specificity for four tests with true sensitivity (SENS) of 0.80, 0.85, 0.90, and 0.95 and specificity (SPEC) of 0.95, 0.90, 0.85, and 0.80 for each of the four tests, respectively.

$SENS_T$	$SPEC_T$	Test	SENS*	SPEC*	$E_{FM}[\log L_{GRE}]$	$E_{FM}[\log L_{FM}]$
1	1	1	0.80	0.95	−1.9015	−1.8953
		2	0.85	0.90		
		3	0.90	0.85		
		4	0.95	0.80		
0.90	0.75	1	0.77	0.94	−2.3313	−2.3279
		2	0.81	0.89		
		3	0.87	0.85		
		4	0.92	0.80		
0.75	0.90	1	0.78	0.94	−2.1905	−2.1872
		2	0.82	0.89		
		3	0.88	0.84		
		4	0.93	0.79		
0.75	0.75	1	0.76	0.94	−2.3633	−2.3606
		2	0.80	0.89		
		3	0.85	0.84		
		4	0.91	0.79		
0.5	0.5	1	0.69	0.89	−2.4734	−2.4733
		2	0.67	0.83		
		3	0.82	0.79		
		4	0.75	0.72		

A Gaussian random effects (GRE) model is assumed, and the true model is a finite mixture (FM) with $\eta_0 = \eta_1 = 0.5$ and $P_d = 0.20$. Asymptotic biases of SENS and SPEC are SENS*−SENS and SPEC*−SPEC, respectively.

Table IV

Simulation results for a common sensitivity and specificity with 5 tests ($J = 5$) under a correctly specified FM model with a sample size of 1000 individuals ($I = 1000$).

$SENS_T$	$SPEC_T$	Model	SENS	SPEC	P_d	$\log L_{FM} \gg \log L_{GRE}$ (per cent)
1	1	FM	0.75 (0.0159)	0.90 (0.005)	0.20 (0.013)	76
		GRE	0.75 (0.016)	0.90 (0.005)	0.20 (0.013)	
1	0.90	FM	0.75 (0.019)	0.90 (0.005)	0.20 (0.014)	66
		GRE	0.74 (0.022)	0.90 (0.005)	0.21 (0.014)	
0.75	0.75	FM	0.75 (0.027)	0.90 (0.007)	0.20 (0.018)	43
		GRE	0.72 (0.039)	0.90 (0.008)	0.21 (0.022)	
0.50	0.50	FM	0.75 (0.038)	0.90 (0.010)	0.20 (0.024)	11
		GRE	0.59 (0.15)	0.90 (0.019)	0.30 (0.11)	

The true model is a finite mixture (FM) model with $\eta_0 = \eta_1 = 0.20$, SENS = 0.75, SPEC = 0.90, and $P_d = 0.20$.

The averages (standard deviations) of estimates are presented, assuming the correctly specified FM model and a misspecified GRE model.

*Values of the log-likelihood for the correctly specified FM model larger by 1 unit than values of the log-likelihood for the incorrect GRE model.

Table V

Simulation examining the predictive values under a simulation identical to the one in Table IV.

$SENS_T$	$SPEC_T$	Model	s					
			0	**1**	**2**	**3**	**4**	**5**
1	1	FM	1.0×10^{-3}	2.2×10^{-2}	0.26	0.84	0.99	1.00
			(3.2×10^{-4})	(5.1×10^{-3})	(3.3×10^{-2})	(2.9×10^{-2})	(4.0×10^{-3})	(1.4×10^{-4})
		GRE	2.5×10^{-3}	2.7×10^{-2}	0.25	0.81	0.98	0.00
			(1.0×10^{-3})	(6.1×10^{-3})	(2.9×10^{-2})	(3.6×10^{-2})	(6.9×10^{-3})	(7.5×10^{-4})
1	0.9	FM	1.0×10^{-3}	2.3×10^{-2}	0.26	0.84	0.99	1.00
			(4.5×10^{-4})	(7.0×10^{-3})	(4.3×10^{-2})	(3.2×10^{-2})	(4.3×10^{-3})	(1.5×10^{-4})
		GRE	5.2×10^{-3}	3.8×10^{-2}	0.28	0.81	0.98	0.00
			(3.3×10^{-3})	(1.3×10^{-2})	(3.9×10^{-2})	(3.9×10^{-2})	(8.0×10^{-3})	(8.4×10^{-4})
0.75	0.75	FM	1.0×10^{-3}	2.3×10^{-2}	0.26	0.84	0.99	1.00
			(6.9×10^{-4})	(1.0×10^{-2})	(7.4×10^{-2})	(5.5×10^{-2})	(6.2×10^{-3})	(2.3×10^{-4})
		GRE	1.1×10^{-2}	5.4×10^{-2}	0.29	0.76	0.97	0.00
			(1.0×10^{-2})	(2.6×10^{-2})	(6.7×10^{-2})	(7.6×10^{-2})	(2.2×10^{-2})	(3.1×10^{-3})
0.50	0.50	FM	1.3×10^{-3}	2.5×10^{-2}	0.27	0.83	0.99	1.00
			(1.2×10^{-3})	(1.7×10^{-2})	(0.12)	(8.6×10^{-2})	(8.6×10^{-3})	(3.5×10^{-4})
		GRE	0.11	0.17	0.45	0.81	0.96	1.00
			(0.11)	(0.12)	(0.22)	(0.21)	(8.3×10^{-2})	(1.7×10^{-2})

We estimate the predictive values $P(d_i=1|\sum_{j=1}^{J} y_{ij}=s)$ for $s = 0, 1, 2, \ldots, J$, where $J = 5$. These values were estimated under the correctly specified (FM) model as well as the misspecified (GRE) model. Averages (standard deviations) of 1000 simulated data sets are presented.

Table VI

Simulation results for a common sensitivity and specificity with 5 tests ($J = 5$) and a sample size of 1000 individuals ($I = 1000$) under a correctly specified finite mixture (FM) model with $\eta_0 = \eta_1 = 0.20$, SENS = 0.75, SPEC = 0.90, and $P_d = 0.50$.

N	SENS	SPEC	P_d
50	0.75 (0.018)	0.90 (0.014)	0.50 (0.026)
100	0.75 (0.015)	0.90 (0.011)	0.50 (0.021)
500	0.75 (0.012)	0.90 (0.008)	0.50 (0.018)
1000	0.75 (0.012)	0.90 (0.008)	0.50 (0.018)
10 000	0.75 (0.011)	0.90 (0.008)	0.50 (0.017)

The imperfect reference test has sensitivity ($SENS_T$) and specificity ($SPEC_T$) both equal to 0.90. $SENS_T$ and $SPEC_T$ are estimated from studies with prevalence 0.50 and different sample sizes (*N*). Averages (standard deviations) of 1000 simulated data sets are presented.

Table VII

Simulation results for four tests with test-specific sensitivity and specificity.

			$X = 1$			$X = 0.90$			$X = 0.75$		
Test	**Est.**	**Truth**	**FM**	**GRE**	**INV**	**FM**	**GRE**	**INV**	**FM**	**GRE**	**INV**
1	SENS	0.80	0.80 (0.018)	0.80 (0.018)	0.80 (0.018)	0.80 (0.020)	0.79 (0.021)	0.80 (0.024)	0.80 (0.026)	0.78 (0.028)	0.80 (0.038)
	SPEC	0.95	0.95 (0.010)	0.95 (0.010)	0.95 (0.010)	0.95 (0.011)	0.94 (0.013)	0.95 (0.017)	0.95 (0.014)	0.93 (0.019)	0.95 (0.034)
2	SENS	0.85	0.85 (0.016)	0.85 (0.016)	0.85 (0.016)	0.85 (0.018)	0.84 (0.018)	0.85 (0.022)	0.80 (0.022)	0.83 (0.025)	0.80 (0.037)
	SPEC	0.90	0.90 (0.014)	0.90 (0.014)	0.90 (0.014)	0.90 (0.016)	0.89 (0.016)	0.90 (0.020)	0.90 (0.019)	0.88 (0.023)	0.90 (0.038)
3	SENS	0.90	0.90 (0.013)	0.90 (0.013)	0.90 (0.013)	0.90 (0.015)	0.89 (0.016)	0.90 (0.020)	0.90 (0.019)	0.88 (0.023)	0.90 (0.037)
	SPEC	0.85	0.85 (0.016)	0.85 (0.016)	0.85 (0.016)	0.85 (0.019)	0.84 (0.020)	0.85 (0.021)	0.85 (0.022)	0.83 (0.025)	0.85 (0.039)
4	SENS	0.95	0.95 (0.010)	0.95 (0.010)	0.95 (0.010)	0.95 (0.011)	0.94 (0.012)	0.95 (0.018)	0.95 (0.014)	0.93 (0.019)	0.95 (0.033)
	SPEC	0.80	0.80 (0.018)	0.80 (0.018)	0.80 (0.018)	0.80 (0.020)	0.79 (0.021)	0.80 (0.023)	0.80 (0.025)	0.78 (0.027)	0.80 (0.039)
$\log L_{FM} > \log L_{GRE}$				96 per cent			91 per cent			88 per cent	

The true model is a finite mixture (FM) model with $\eta 0 = \eta 1 = P_d = 0.50$ with a sample size of 1000 ($I = 1000$). Models are fit assuming the correctly specified FM model, the incorrect GRE model, and the approach which uses each test individually (IND). It is assumed that $SENS_T$ and $SPEC_T$ are known with values $SENS_T = SPEC_T = X$, where $X = 1$, 0.90, and 0.75. Average estimates (standard deviations) from 1000 simulated data sets are presented.