On Estimating Diagnostic Accuracy From Studies With Multiple Raters and Partial Gold Standard Evaluation

Paul S. Albert* and Lori E. Dodd
Biometric Research Branch
Division of Cancer Treatment and Diagnosis
National Cancer Institute
March 15, 2006

* *email:Albertp@ctep.nci.nih.gov*

KEY WORDS: Diagnostic error; Misclassification; Latent class models; Semi-latent class models

Summary

We are often interested in estimating sensitivity and specificity of a group of raters or a set of new diagnostic tests in situations in which gold standard evaluation is expensive or invasive. Numerous authors have proposed latent modeling approaches for estimating diagnostic error without a gold standard. Albert and Dodd (2004) showed that, when modeling without a gold standard, estimates of diagnostic error can be biased when the dependence structure between tests is misspecified. In addition, they showed that choosing between different models for this dependence structure is difficult in most practical situations. While these results caution against using these latent class models, the difficulties of obtaining gold standard verification remain a practical reality. We extend two classes of models to provide a compromise that collects gold standard information on a subset of subjects but incorporates information from both the verified and non-verified subjects during estimation. We examine the robustness of diagnostic error estimation with this approach and show that choosing between competing models is easier in this context. In our analytic work and simulations, we consider situations in which verification is completely at random as well as settings in which the probability of verification depends on the actual test results. We apply our methodological work to a study designed to estimate the diagnostic error of digital radiography for gastric cancer.

1 Introduction

Diagnostic and screening tests are important tools of modern clinical decision making. These tests help to diagnose illness to initiate treatment (e.g., a throat culture for streptococcal infection) or to identify individuals requiring more extensive follow-up (e.g., mammography screening for breast cancer). Estimation of sensitivity and specificity, measures of diagnostic accuracy, requires knowledge of the true disease state, which is assessed by a gold or reference standard. (Throughout, we use both "gold standard" and "reference standard" to mean the accepted standard for diagnosis). Gold standard evaluation may be expensive, time-consuming or unethical to perform on all subjects, and is commonly difficult to obtain in clinical studies. Latent-class models offer a tempting alternative because assessment of the true status is not necessary. However, it has been shown that latent class models for estimating diagnostic error and prevalence may be problematic in many practical situations (Albert and Dodd, 2004). Specifically, they showed that, with a small number of tests, estimates of diagnostic error were biased under a misspecified dependence structure, yet in many practical situations it was nearly impossible to distinguish between models based on the observed data. The lack of robustness of these models is concerning; however, the limitations of obtaining gold standard is a practical reality and reasonable alternatives are desirable.

Although it may be very difficult to obtain the gold standard on all subjects, in many cases, it may be feasible to obtain gold standard information on a fraction of subjects (a.k.a, partial gold standard evaluation). In radiological studies, for example, gold standard evaluation usually requires multiple radiologists simultaneously examining images and clinical information. This may be an infeasible proposition for many studies to collect the gold standard on all subjects. However, it may be very feasible to obtain gold standard information on a fraction of study subjects. Thus, methodological approaches which incorporating partial gold standard information may be an attractive alternative to latent class modeling.

Our application is a medical imaging study to compare conventional versus digital radiography for diagnosing gastric cancer (Iinuma, et al., 2000). In this study, six radiologists evaluated 225 images on either conventional (n=112) or digital (n=113) radiography, to compare the sensitivity and specificity across techniques and radiologists. A gold standard evaluation was obtained from three independent radiologists simultaneously reviewing clinical information along with all imaging data to provide a reference truth evaluation of the image. Specifically, these radiologists reviewed clinical information such as patient characteristics, chief symptoms, purposes of the examination, endoscopic features, and histologic findings in biopsy specimens. This time-consuming consensus review was done on all 225 images, although it may not be feasible in larger studies or in other studies with more limited resources. Rater-specific as well as over-all sensitivity and specificity were estimated by treating the consensus review by the three independent radiologists as gold standard truth. Our methodological development will focus on the data from this study.

Although our primary example is in radiology, the problem occurs more generally in medicine. For example, similar problems exist for the evaluation of biomarkers in which one wishes to compare the diagnostic accuracy of a series of tests, where a gold standard exists, but is very expensive. See, for example Van Dyck, et al. (2004), in which a set of tests for herpes simplex virus type 2 (HSV-2) were compared, but only a subset of samples were verified with the reference standard Western blot.

In this paper, we extend two classes of models, originally proposed for modeling diagnostic error on multiple tests without a gold standard (Albert and Dodd, 2004), to the situation of estimating diagnostic error for a partially verified design. We examine the robustness of these models to the assumed dependence structure between tests. In particular, we examine bias and model selection using asymptotic results and simulation studies. We examine whether observing gold standard information on a small percentage of cases improves the lack of robustness to assumptions on the dependence between tests found when modeling without a gold standard. In section 2, we describe our approach, which considers various models for the dependence between tests. In Section 3, we fit the various classes of models to the gastric cancer dataset and show that the results are quite different when we use the reference standard evaluation or when we model without a gold standard. In Section 4, we investigate the asymptotic bias from misspecifying the dependence structure under full as well as partial reference sample evaluation. Simulations examining the finite sample properties of partial reference sample verification are described in Section 5. We illustrate the effect of partial reference sample verification using the gastric cancer dataset in Section 6. A discussion follows in Section 7 in which we make general recommendations.

2 Models

Let $\boldsymbol{Y}_i = (Y_{1i}, Y_{i2}, ..., Y_{iJ})'$ be dichotomous test results for individual i ($i = 1, 2, 3, ..., I$), with Y_{ij} denoting the result from the jth of J tests. We denote d_i as the true unobserved disease status for patient i and v_i as an indicator of whether the ith patient is verified by a reference standard ($v_i = 1$ if verified and $v_i = 0$, otherwise). When a patient is verified, the contribution to the likelihood is $L_i = P(\boldsymbol{Y}_i)P(d_i)P(v_i = 1|\boldsymbol{Y}_i)$. Similarly, when a patient is not verified, the contribution is $L_i = P(v_i|\boldsymbol{Y}_i)\sum_{l=0}^{1} P(\boldsymbol{Y}_i|d_i = 1)P(d_i = 1)$. In a general form, when the verification mechanism does not share parameters with the probability of disease $P(d_i)$ or the diagnostic accuracy $P(\boldsymbol{Y}_i|d_i)$, then the contribution of the ith patient to the likelihood (L_i) is proportional to:

$$L_i \propto [P(\boldsymbol{Y}_i|d_i)P(d_i)]^{v_i}[\sum_{l=0}^{1} P(\boldsymbol{Y}_i|d_i = l)P(d_i = l)]^{1-v_i}, \tag{1}$$

where $P(d_i = 1)$ is the disease prevalence which will be denoted by P_d.

There are three types of verification processes. First, consider verification that is completely at random, which occurs if the verification process is a simple random sample chosen independently from the test results $\boldsymbol{Y}_i$. The proportion of individuals verified is denoted r, where $r = P(v_i = 1)$. Second, consider verification in which the probability of verification depends on $\boldsymbol{Y}_i$, which we denote as $r_{\boldsymbol{Y}_i} = P(v_i = 1|\boldsymbol{Y}_i)$. Of particular interest is when the probability of verification depends on the number of positive tests, $r_s = P(v_i = 1|\sum_{j=1}^{J} y_{ij}),\ s = 1, 2, ..., J$. This type of verification has been referred to as verification biased sampling (Pepe, 2003). An important special case, called extreme verification biased sampling occurs when the gold standard test is obtained only on test positive subjects because it requires an invasive procedure such as surgery, which is unethical to perform on all subjects if the experimental tests $\boldsymbol{Y}_i$ are negative. Various authors have proposed models for analyzing the results of two tests under extreme biased sampling (Walter, 1999; Hoenig et al., 2002; Merwe and Manitz, 2002). The third type of verification occurs when the probability of verification depends on the true disease status, which is only known for those patients verified, denoted $r_{d_i} = P(v_i = 1|d_i)$. This so called, non-ignorable verification has been discussed by various authors including Kosinski and Barnhart (2003) and Baker (1995). We focus only on the first two types of verification processes.

We consider two different ways to specify $P(\boldsymbol{Y}_i|d_i)$ that were originally developed for estimating diagnostic error without a gold standard. The Gaussian random effects and finite mixture models are very different formulations for describing conditional dependence between tests which both have attractive features. The Gaussian random effects (GRE) model (Qu et al., 1996) introduces dependence across tests by assuming that $(Y_{ij}|d_i, b_i)$ are independent Bernoulli with proportion given by $\Phi(\beta_{jd_i} + \sigma_{d_i} b_i)$, where the random variables b_i are standard normal and Φ is the cumulative distribution function of a standard normal distribution. Under this model, $P(\boldsymbol{Y}_i|d_i) = \int \{\prod_{j=1}^{J} P(Y_{ij}|d_i, b)\}\phi(b)db$, where $\phi(b)$ is the standard normal density. Under the GRE model, the sensitivity and specificity of the jth test is given by $\Phi(\beta_{j1}/\sqrt{1+\sigma_1^2})$ and $1 - \Phi(\beta_{j0}/\sqrt{1+\sigma_0^2})$, respectively. A substantially different model for incorporating dependence is the finite mixture (FM) model (Albert et al., 2001; Albert and Dodd, 2004) 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. Let l_{id_i} be an indicator of whether the i^{th} subject, given disease status d_i, is always classified correctly, so that $l_{i1} = 1$ when a true positive subject is always positive and $l_{i0} = 1$ when a truly negative is always rated negative. Further, define $\eta_0 = P(l_{i0} = 1)$ and $\eta_1 = P(l_{i1} = 1)$. Test results Y_{ij} given d_i and l_{id_i} are independent Bernoulli with probability

$$P(Y_{ij} = 1|d_i, l_{id_i}) = \begin{cases} 1 & \text{if } d_i = 1 \text{ and } l_{i1} = 1 \\ 0 & \text{if } d_i = 0 \text{ and } l_{i0} = 1 \\ \omega_j(1) & \text{if } d_i = 1 \text{ and } l_{i1} = 0 \\ 1 - \omega_j(0) & \text{if } d_i = 0 \text{ and } l_{i0} = 0, \end{cases} \tag{2}$$

where $\omega_j(d_i)$ is the probability of the j^{th} test making a correct diagnosis when the individual is subject to diagnostic error ($l_{i1} = 0$ or $l_{i0} = 0$). Under the finite mixture model, the sensitivity and specificity of the jth test are $\eta_1 + (1 - \eta_1)\omega_j(1)$ and $\eta_0 + (1 - \eta_0)\omega_j(0)$, respectively. Under both the GRE and FM models, estimates of a common sensitivity and specificity across J tests can be obtained by assuming $\beta_{0l} = \beta_{1l} = ... = \beta_{Jl} = \beta_1$ and $\omega_1(l) = \omega_2(l) = ... = \omega_J(l) = \omega(l)$, for $l = 0, 1$.

Depending on the application, the FM or the GRE model may better describe the dependence structure between tests. Both models need to be compared with a simple alternative which is nested within both of these conditional dependence models. The conditional independence (CI) model which assumes the tests are independent given the true disease status provides such an alternative. The GRE model reduces to the CI model when $\sigma_0 = \sigma_1 = 0$, while the FM model reduces to the CI model when $\eta_0 = \eta_1 = 0$.

For each of the models, estimation is based on maximizing $L = \prod_{i=1}^{I} L_i$, where L_i is given by (1). Standard errors can be estimated with the Bootstrap (Efron and Tibshirani, 1993).

3 Analysis of Gastric Cancer Data

We estimate prevalence, sensitivity and specificity of digital radiography for gastric cancer using the likelihood in (1) and the GRE, FM, and CI models, under both complete and no verification. Table 1 shows the overall estimates of prevalence, sensitivity and specificity for digital radiography with the consensus measurements as a gold standard and with no gold standard. Estimates were obtained by assuming a common sensitivity and specificity across the 6 raters, and were derived under the CI, as well as the GRE and FM models. Bootstrap standard errors are also presented under each model. Interestingly, under complete verification, overall estimates of prevalence, sensitivity and specificity as well as their bootstrap standard errors were nearly identical across the three classes of models. In addition, these estimates were identical to estimates obtained by Iinuma et al. (2000) using generalized estimating equations (Liang and Zeger, 1986), a procedure known to be insensitive to assumptions on the dependence structure between tests. These results suggest that estimates of prevalence, sensitivity, and specificity are insensitive to the dependence structure between tests under complete verification. When no gold standard information is incorporated, estimates of prevalence and diagnostic error differ across models for the dependence between tests. This is consistent with results by Albert and Dodd (2004) who showed that diagnostic error estimation may be sensitive to assumptions on the dependence between tests when no verification is performed.

By the likelihood principle, we compare models based on a comparison of the likelihood values. Using the gold standard, the log-likelihoods were -314.63, -300.36, and -305.45, for the CI, GRE, and FM models, respectively (there are 3, 5, and 5 parameters for each model, respectively). We compared the GRE and FM models with the CI model using a likelihood ratio test since the CI model is nested within both of these conditional dependence models. Since the parameters which characterize the conditional dependence are on the boundary ($\sigma_0 = \sigma_1 = 0$ for the GRE model and $\eta_0 = \eta_1 = 0$ for the FM model) under the null hypothesis corresponding to a CI model, the standard likelihood ratio theory is inappropriate (Self and Liang, 1997). We conducted a simulation study to obtain the reference distribution under the null hypothesis by simulating 10,000 datasets under the estimated CI model and evaluating the likelihood ratio test of $\sigma_0 = \sigma_1$ and $\eta_0 = \eta_1 = 0$ corresponding to the GRE model and FM models. Based on the above observed log-likelihoods and the simulated reference distribution, we reject the independence model in favor of the GRE and FM models ($P < 0.001$, for both models). Further, parameter estimates characterizing the conditional dependence under both conditional dependence models are sizable. For the GRE model, $\widehat{\sigma}_0 = 1.1$ and $\widehat{\sigma}_1 = 0.37$ and for the FM model $\widehat{\eta}_0 = 0.31$ and $\widehat{\eta}_1 = 0.38$, respectively. A comparison of the two non-nested GRE and FM models can be made by directly comparing the two log-likelihoods since both models have the same number of parameters. Under complete gold standard evaluation, this comparison clearly favors the GRE model.

For the no gold standard case, the log-likelihoods for the CI, GRE, and FM models were -283.19, -280.16, and -280.30, respectively. Consistent with Albert and Dodd (2004), these results suggest that, although it is easy to distinguish between conditional dependence and a conditional independence model (likelihood ratio tests computed as described above for complete verification showed evidence for conditional dependence; P-values for the comparisons of the GRE and FM models relative to the CI model were 0.009 and 0.016, respectively), it may be difficult to choose between the two models for conditional dependence with no gold standard.

Table 2 shows rater-specific estimates of sensitivity and specificity, along with prevalence for models which incorporate the gold standard information and those that do not. As with the overall estimates of sensitivity and specificity, individual rater estimates are nearly identical across models for the dependence between tests as well as to the rater-specific estimates presented in Iinuma et al. (2000). In contrast, estimates obtained using no gold standard information were highly model dependent and were very different from those estimates which used the gold standard information.

Thus, modeling approaches with complete verification appear to be more robust against misspecification of the dependence structure between tests, while approaches with no verification appear to lack robustness. A natural question is how the statistical properties of the estimation improve with an increasing proportion of gold standard evaluation. This will be the primary focus of this paper. We discuss asymptotic and simulation results before returning to this example and varying the amount of verification. We focus on comparing the GRE and FM models since it has been shown in Albert and Dodd (2004) that it is difficult to distinguish between these rather different models with no gold standard evaluation.

4 Asymptotic Results

We examined the asymptotic bias when the dependence structure is misspecified as a function of the proportion of samples receiving gold standard evaluation. For simplicity, we examine this bias for the case when interest focuses on estimating a common sensitivity and specificity across raters (denoted as $SENS$ and $SPEC$, respectively). We examined both verification that is completely at random and verification biased sampling. The misspecified maximum-likelihood estimator for the model parameters, denoted by $\widehat{\boldsymbol{\theta}}^*$, converges to the value $\boldsymbol{\theta}^*$, where

$$\boldsymbol{\theta}^{*} = \arg\,\max_{\boldsymbol{\theta}} E_T[\log L(\boldsymbol{Y}_i, \boldsymbol{\theta})], \tag{3}$$

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

$$E_T(\log L_M) = E_T[\log L(\boldsymbol{Y}_i, \boldsymbol{\theta})]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{*}} \tag{4}$$

denotes the expectation (taken under the true model T) 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 and specificity converge to $SENS^{*}$ and $SPEC^{*}$ 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 bias for sensitivity and specificity is defined as $SENS^{*} - SENS$ and $SPEC^{*} - SPEC$, respectively.

First, we examined the case of completely at random verification (i.e., $r_s = r$ for all $s = 1, 2, .., J$). We initially examined the asymptotic bias of estimators of sensitivity and specificity when we falsely assumed a GRE model and when the true model is a FM model as well as when we falsely assumed a FM model and the true model is a GRE model. dependence structure. This reciprocal misspecification with the FM and GRE models is an extreme type of misspecification since the two models are so different.

Table 3 shows the results for various proportions of completely-at-random verification for five tests and a presumed constant sensitivity and specificity, with the true model being the FM model and the misspecified model being the GRE model. When we have no gold standard information ($r = 0$), there is serious bias under a misspecified dependence structure and the expected individual contribution to the log-likelihood under the correctly specified model is nearly identical (to more than 6 digits) to the expected log-likelihood under the correctly specified model, which is consistent with results reported in Albert and Dodd (2004). Thus, with no gold standard reference and with five tests, estimates of diagnostic error may be biased under a misspecified dependence structure, yet it may be very difficult to distinguish between models in most situations. As little as 2% gold standard verification ($r = 0.02$), reduces the bias considerably and the expected log-likelihoods are no longer nearly identical making it simpler to distinguish between models. With 20% verification, the bias is small. For complete verification (r=1), marginal quantities such as sensitivity and specificity are nearly unbiased under a misspecified dependence structure. This is consistent with work by Tan et al. (1999) and Heagerty and Kurland (2001) who showed for clustered binary data that marginal quantities (which sensitivity, specificity, and prevalence are) are robust to misspecification of the dependence structure. The large differences in expected log-likelihoods suggest that it will be relatively simple to distinguish between models.

Table 4 shows asymptotic bias with five tests when the true model is the GRE model and the misspecified model is the FM model. As in Table 3, there is substantial asymptotic bias under the misspecified model when there is no gold standard evaluation. In addition, the expected log-likelihood for the misspecified model is nearly identical to the expected log-likelihood for the correctly specified model, again showing the difficulty in choosing between competing models with no gold standard information with few tests. Similar to the results in Table 3, estimates of prevalence, sensitivity, and specificity are asymptotically unbiased under the misspecified model when there is complete gold standard evaluation ($r = 1$). Unlike the results in Table 3, a larger percentage of verification (about 50%) is necessary to achieve approximate unbiasedness. In both cases, however, a small percentage of verification results in different expected log-likelihoods under the true and misspecified models, suggesting that it is simpler to choose between competing models with even a small percentage of gold standard verification.

Tables 3 and 4 provide an assessment of asymptotic bias under reciprocal model misspecification for both the FM and GRE models when the sensitivity and specificity are 0.75 and 0.9, respectively. We also examined the relative asymptotic bias for a wide range of sensitivity and specificity (a grid ranging from values of 0.65 to 0.95 for both sensitivity and specificity) corresponding to the cases specified in these tables for $r = 0.5$. Figure 1 shows the results corresponding to the models and parameters described in Table 3 when $\sigma_1 = 3$. Over the wide range of sensitivity and specificity, the maximum relative % bias for sensitivity was 2.8% and for specificity was 5.1%. Other scenarios provided similar results with all percent biases being less than 6% over the grid (data not shown).

We examined asymptotic bias of the FM and GRE models under alternative dependence structures. Specifically, we examined the asymptotic bias when the true conditional dependence structure $P(\boldsymbol{Y}_i|d_i)$ is a Bahadur model (Bahadur, 1961), a log-linear model (Cox, 1972), and a Beta-binomial model, where a description of each of these models is provided in the Appendix B. All three alternative models were formulated so they had the same number of parameters as the GRE and FM models. For the Bahadur model, we considered the special case of only pairwise conditional dependence between tests (i.e., all three and higher-order interactions are set to zero). For example, the conditional distribution of $\boldsymbol{Y}_i|d_i$ is $P(\boldsymbol{Y_i}|d_i = 1) = \big\{ \prod_{j=1}^{J} SENS^{Y_{ij}}(1 - SENS)^{1-Y_{ij}} \big\}(1 + \sum_{j<k} \rho_1 e_{ij}e_{ik})$, where $e_{ij} = \frac{Y_{ij}-SENS}{\sqrt{SENS(1-SENS)}}$ and $\rho_1 = E[e_{ij}e_{ik}|d_i = 1]$ for all $j \neq k$ for any i. Similar to Tables 3 and 4 for reciprocal model misspecification, we evaluated asymptotic bias of sensitivity and specificity for an increasing fraction r of completely at random verification under a GRE and FM model when the true model was the Bahadur model. For five tests ($J = 5$), $SENS = SPEC = 0.75$, and $P_d = 0.20$, sensitivity and specificity were nearly asymptotically unbiased under both the GRE and FM models with 20% completely random verification. For example, under a GRE model, $SENS^*$=0.50, 0.63, 0.72, 0.74, and 0.75 for r=0, 0.02, 0.2, 0.5, and 1. Under a FM model, $SENS^*$=0.61, 0.73, 0.78, 0.76, and 0.75 for these values of r.

For all three alternative models, we examined the bias in sensitivity and specificity of the GRE and FM models with 50% completely random verification over a wide range of sensitivity and specificity values (identical to the grid described for Figure 1) for a prevalence of 0.20. Table 5 shows that the maximum relative asymptotic bias was less than 7% for both sensitivity and specificity for all three alternative models. Thus, estimates of diagnostic error appear to be quite robust with 50% completely random verification. When prevalence was very low or very high (e.g., below 5% or above 95%) there was more substantial bias under certain model misspecification with 50% completely random verification. For example, for a prevalence of 0.05 when the true model was the log-linear model, there was a maximum bias of 10% under a GRE model (as compared a maximum bias of 4.3% for a prevalence of 0.20). However, unlike when there is no gold standard evaluation ($r = 0$), it is much easier to identify the better fitting model using likelihood and other criterion for model assessment. Further, for a rare disease, completely random verification would not generally be recommended due to efficiency considerations.

While random verification is of concern for our application, we also consider verification biased sampling because it is so common. We examine asymptotic properties under a misspecified dependence structure with verification biased sampling. Table 6 shows asymptotic bias and expected log-likelihoods for the situation in which a random sample of cases *among those who test positive on at least one of the 5 tests* are verified (e.g., extreme verification biased sampling) and where the true model is the FM model and the misspecified model is the GRE model. Interestingly, these results suggest that, in some cases, an increase in the proportion verified can result in an increase in bias under the misspecified model. For example, when $\eta_1 = 0.5$ and $\eta_0 = 0.2$, the estimator of sensitivity is only slightly asymptotically biased ($SENS^*$=0.77) with no gold standard evaluation ($r_s = 0$ for $s = 0, 1, 2, .., 5$) and substantial bias ($SENS^*$=0.57) under complete verification of any case with positive tests ($r_0 = 0$ and $r_s = 1$ for $s = 1, 2, .., 5$). This result is consistent with our simulation results, which are presented in the next section. This problem occurs more generally under a wide range of verification biased sampling. For example, situations where one over-samples discrepant cases can result in bias under model misspecification. Bias can also increase with an increasing proportion of verification of discrepant cases. As an illustration, under completely random verification, when the true model is the FM model as described in Table 3 with $\eta_1 = 0.5$, the sensitivity converges to $SENS^* = 0.76$ and is nearly unbiased when $r = 0.2$. When we over-sample discrepant cases $r_0 = r_5 = 0.20$ and $r_s = 0.4$, for $s = 1, 2, 3$, and 4, estimates of sensitivity are more asymptotically biased ($SENS^* = 0.73$). The asymptotic bias is increased further ($SENS^* = 0.71$) when r_s, $s = 2, 3$, and 4 is changed from 0.4 to 1. We found similar results when the true model was the GRE model and the misspecified model was the FM model.

In the next section, we examine the finite sample results for both robustness and efficiency when we observe partial gold standard information.

5 Finite Sample Results

We examine bias, variability and model selection of the different models using simulation studies. Table 7 shows the effect of model misspecification on estimates of prevalence, sensitivity and specificity when the true model is a FM model and we fit the misspecified GRE. Results are shown for sample sizes of $I = 100$ and $I = 1000$ and for various proportions of random verification r. Similar to simulations in Albert and Dodd (2004), we found that when $r = 0$ estimates of sensitivity, specificity, and prevalence are biased under a misspecified model and it is difficult to distinguish between models based on likelihood comparisons. In addition, estimates under the misspecified GRE model are substantially more variable than estimates under the correctly specified FM model. However, with only a small percentage of samples verified, estimation of sensitivity, specificity, and prevalence have improved statistical properties. Table 7 shows that bias is substantially reduced when only 5% of cases are verified. With as little as 20% random verification, estimates of sensitivity, specificity, and prevalence are nearly unbiased under model misspecification. In addition, variance estimates are very similar under the misspecified model relative to the correctly specified model. Under complete verification ($r = 1$), there is essentially no effect to misspecifying the dependence structure. The table suggests that there are other advantages to measuring the gold standard test on at least a fraction of samples or individuals. First, there is a large pay-off in efficiency. For sensitivity, under the correct FM model with $I = 1000$, the efficiency gain relative to no gold standard information ($r = 0$) is 46%, 276%, and 640% for 5%, 20%, and 100% gold standard evaluation (these calculations were based on variance estimates computed to the fourth decimal place, while the standard errors in Table 7 are only presented to the second decimal place). This decrease in variance is even more sizable under the misspecified GRE model. Second, it becomes increasingly easier to distinguish between models for the dependence structure with increasing r. In Table 7 we show the percentage of times the correctly specified FM model is chosen to be superior than the misspecified GRE based on the criterion of a separation of likelihoods greater than 1. With five tests ($J = 5$) and a sample size of $I = 1000$, the correctly specified FM was declared to be superior 12% of the cases when there was no gold standard tests. The ability to choose the correct model increased dramatically with even a small fraction of gold standard evaluation. With only 5%, 20%, and 100% verification, the correct model was identified in 45%, 64%, and 79% of the cases.

Table 8 shows the effect of model misspecification on estimates of sensitivity and specificity when the true model is a GRE model and the misspecified model is the FM model. As with the asymptotic results in this situation, a random sample of larger than 20% reference standard evaluation is needed to get approximately unbiased estimates under the misspecified model. However, unlike when $r = 0$, where it is difficult to choose the correct model (by the criterion that the log-likelihood for the GRE model was larger than the log-likelihood for the FM model by more than one), we can choose between the GRE and FM model with high probability when $r = 0.2$.

We also examined the robustness of the GRE and FM models when the true dependence structure is governed by a Bahadur model. Specifically, we simulated data with the conditional dependence structure ($P(\boldsymbol{Y}_i|d_i)$) given by a Bahadur model with pairwise correlation of 0.2 and all three and higher way correlations equal to zero. Further, data were simulated corresponding to $SENS = SPEC = 0.75$, $P_d = 0.20$, $I = 1000$, and $J = 5$. For $r = 0$, there was substantial bias under the misspecified GRE and FM models. In this case, the average sensitivity and specificity were 0.52 (SE=0.09) and 0.68 (SE=0.05) under the GRE model and 0.65 (SE=0.06) and 0.87 (SE=0.04) under the FM model. In addition, it was difficult to distinguish between the correctly specified Bahadur model and the GRE or FM models. For example, the misspecified FM model had a larger likelihood than the correctly specified

Bahadur model in 40% of the simulated realizations. Both the GRE and FM models resulted in nearly unbiased estimates of sensitivity and specificity for $r = 0.20$. The average estimates of sensitivity and specificity were 0.73 (0.03) and 0.74 (0.01) for the GRE model and 0.77 (0.03) and 0.77 (0.01) for the FM model when r=0.2. Also, the likelihood for the correctly specified model was larger than the likelihood of the FM and GRE model in greater than 99% of the simulated realizations. Thus, with only 20% completely random verification, both the GRE and FM models are robust to model misspecification and it is relatively easy to distinguish between models.

Table 9 shows simulation results for the case of four raters under a correctly specified FM model and misspecified model GRE model. Estimates of sensitivity and specificity, which are seriously biased with no gold standard evaluation are nearly unbiased under the misspecified model with 20% random verification. As in Table 7, this table illustrates the pay-off in efficiency with at least some partial gold standard evaluation, under both the correct and misspecified model. This table also shows the percentage of realizations where the FM model has a larger likelihood than the GRE. Unlike with no gold standard evaluation, the FM model is almost always correctly identified with 20% verification. In addition, unlike with $r = 0$, models with 20% verification result in the correct ordering of sensitivity and specificity almost all the time. We also performed simulations for the case of four raters when the true model is a GRE model and the misspecified model is the FM model. Under the misspecified FM model, bias is substantially reduced for $r = 0.20$ as compared to $r = 0$. Furthermore, estimates of sensitivity, specificity, and prevalence computed under the FM model were nearly unbiased for $r = 0.5$ (data not shown).

Next, we examine verification biased sampling. Our asymptotic results show that estimates of diagnostic error and prevalence can be biased when we over-sample discrepant cases under a misspecified model, which was in contrast to results with random verification. We conducted simulations to examine this further. We examine bias in sensitivity, specificity, and prevalence estimates from a GRE model when the FM model is the correct model. We simulated under a FM model with $J = 5$, $I = 100$, $\eta_0 = 0.20$, $\eta_1 = 0.50$, $P_d = 0.20$,

$SENS = 0.75$, and $SPEC = 0.90$ (same parameters as in rows 5-8 in Table 3) and fit both the correctly specified FM and the misspecified GRE model. When all individuals with at least one positive value were verified ($r_s = 1$ for $s = 1, 2, .., 5$ and $r_0 = 0$), we had sizable bias under the misspecified model. Average estimates of prevalence, sensitivity, and specificity were 0.27 (SE=0.08), 0.61 (SE=0.13), and 0.89 (0.02) under the misspecified GRE model and were 0.20 (0.04), 0.75 (0.07), and 0.90 (0.02) under the correct model.

Under the correctly specified model, over-sampling discrepant cases may improve precision of our estimates. Thus, an interesting question is whether the increase in efficiency from over-sampling discrepant cases is worth the potential of serious bias under a misspecified model. We conducted a simulation where we simulated under a finite mixture model and fit both the correctly specified FM models and the misspecified GRE model both under completely random verification and under a verification process where we over-sample discrepant cases. We simulated data with $J = 5$, $I = 1000$, $P_d = 0.2$, $\eta_1 = 0.5$, $\eta_0 = 0.2$, $SENS = 0.75$ and $SPEC = 0.90$. We over sampled by obtaining a gold standard result on 40% of discrepant cases and only 5% of cases where $\boldsymbol{Y}_i$ are all concordant. For the completely at random verification cases, we chose 21% verification to correspond to the overall proportion of verification in the over sampling cases. Figure 2 shows the distribution of sensitivity estimates for each of the four scenarios. The figure shows that there is an efficiency gain in estimating sensitivity by over-sampling discrepant cases. Specifically, there is a 28% efficiency gain in over-sampling as compared to completely random verification. In addition, the figure demonstrates the robustness of sensitivity estimates to model misspecification under completely random verification and the lack of robustness under over-sampling. In this particular case, the pay-off in efficiency with over-sampling under the correct model is small relative to the potential for bias due to model misspecification. Furthermore, the correct model was definitely selected more often under completely random verification than under over-sampling. The FM model had a likelihood greater than 1 more than the GRE model in over 99.5% and 54% of the simulations under completely random verification and under the mechanism that over-samples discrepant cases, respectively.

6 Gastric Cancer Example Continued

Next we return to the gastric cancer data set and use only partial gold standard evaluation. Our initial focus is on examining verification which is completely at random. We evaluated designs with different probabilities of verification (r). In order to capture the variability associated with different amounts of verified sampling, we resample data with replacement and incorporate the reference standard on a given image with probability r. Table 10 shows results for an assumed common and for an assumed rater-specific sensitivity and specificity for r ranging from 0.1 to 0.8. In each situation, we fit both the FM and the GRE models. A comparison of these results with those presented for complete verification and for no gold standard evaluation (Tables 1 & 2) is most revealing. The results suggest that the common as well as the rater-specific estimates for $r = 0.50$ are close to those presented for complete verification. In addition, the results for $r = 0.2$, although not very close to those presented for the complete verification case, are substantially closer than those estimated with the latent class models under $r = 0$ (Table 2).

We also examined extreme bias verification. Specifically, we evaluated a design whereby we verified all images in which at least one of the 6 radiologists rated the image positive for gastric cancer (52% of images were declared positive by at least one radiologist). As with random verification, we constructed data sets by re-sampling images with replacement and incorporating reference standard information whenever a positive image for any radiologist was recorded. For a common sensitivity and specificity, estimates of sensitivity, specificity, and prevalence were 0.78 (SE=0.05), 0.90 (0.01), and 0.23 (0.04) for the FM model and 0.72 (0.11), 0.90 (0.02), and 0.23 (0.06) for the GRE model, respectively. There was greater discrepancy between the estimates across the two models under extreme verification bias than for a comparable proportion verified under a completely random verification mechanism (r=0.50 in Table 10). Large differences between the FM and the GRE model for rater-specific estimates were also found (data not shown). These results, along with the analytic and simulation results, demonstrate less robustness under verification biased sampling.

7 Discussion

It has been shown in previous work that estimates of diagnostic error and prevalence are biased under a misspecified model for the dependence between tests and that, with only a small number of tests, it is difficult to distinguish between models for the dependence structure using likelihood and other model diagnostics (Albert and Dodd, 2004). Under complete verification, results on generalized linear mixed models would suggest that the estimation of marginal quantities (which prevalence, sensitivity, and specificity are) are insensitive to misspecification of the dependence between tests (Tan et al., 1999; Heagerty and Kurland, 2001). Our results confirm this. Furthermore, we showed that it is much simpler to distinguish between models with complete verification. A natural question is whether gold standard verification on even a small percentage of cases improves the statistical properties of estimators of sensitivity, specificity, and prevalence. We examined both whether observing partial verification lessens the bias when the dependence structure is misspecified and whether one is able to more easily distinguish between different models for the dependence structure between tests. For the situation where verification is independent of the test results $\boldsymbol{Y}_i$, gold standard evaluation on even a small percentage of cases greatly lowers the bias for estimating prevalence, sensitivity, and specificity, under a misspecified model. In addition, identifying the correct model for the dependence structure using likelihood comparisons becomes much easier with even a small percentage of gold standard evaluation. Although there are advantages to performing the gold standard test on as many individuals as possible, this is not often possible due to limited resources. Our results suggest that between 20% and 50% gold standard evaluation results in large improvements in robustness, efficiency, and the ability to choose between competing models over no gold standard information. If the gold standard test is expensive, performing the gold standard test on more than 50% of patients may not be cost-effective.

We also examined situations in which the probability of verification depends on observed test results (i.e., verification biased sampling). An important special case of verification biased sampling is extreme verification biased sampling where individuals who test negative on all tests do not receive gold standard evaluation. Such verification sampling occurs in situations where the gold standard is invasive (e.g., surgical biopsy) and it is considered unethical to subject a patient to the invasive test when there is little evidence for disease. Unlike for a single test where sensitivity, specificity, and prevalence are not identifiable under extreme verification bias sampling (Begg and Greenes, 1983; Pepe, 2003), these quantities are identifiable with multiple tests and an assumed model for the dependence between these tests. However, unlike the case where verification is completely at random, estimates of sensitivity, specificity, and prevalence may not be robust to misspecification of the dependence between tests with a large fraction of verification.

A gold standard can be defined in various ways depending on the scientific interest. The gold standard test could be a laboratory test, a consensus evaluation of an image, or an assessment of clinical disease. The nature of the gold standard will determine how diagnostic accuracy is interpreted. In the gastric cancer study, the gold standard was a consensus assessment (across three radiologists) of all available clinical information including imaging data. All suspect gastric cancers were confirmed with biopsies, while patients who were negative had limited follow-up of two months to see if gastric cancer symptoms developed. A longer follow-up would have been ideal in assuring that these negative cases did not develop gastric cancer.

Other types of verification biased sampling schemes may be employed to improve efficiency. For example, our simulation results show that over-sampling discrepant cases can result in improved efficiency over sampling completely at random. Our results further show that, although over-sampling discrepant cases can improve efficiency, such a strategy loses the attractive feature of decreasing bias with an increasing proportion of verification found for a completely random verification mechanism. In addition, our results suggest that for a comparable proportion of verification, choosing the correct model for the dependence between tests is more difficult for a verification process in which we over-sample discrepant cases as compared with completely random verification.

Irwig, et al. (1994) and Tosteson (1994) have considered optimal design strategies for the case of a single diagnostic test. Optimal design for multiple correlated tests is an area for future research. However, the choice of an optimal design will depend heavily on assumed models and parameter values for the dependence between tests. For this reason, we question the practicality of developing an optimal design in this situation.

A common criticism of latent class models for estimating sensitivity, specificity, and prevalence without a gold standard is that, without a gold standard, it is difficult to conceptualize sensitivity and specificity (Alonzo and Pepe, 2003). Partial verification lessens the problem of conceptualizing the truth since a gold standard test needs to be defined and evaluated on at least a fraction of the cases.

The different models presented for analyzing partial verification data use a latent class structure for observations that do not have gold standard evaluation. In contrast with the full latent class modeling used when there is no gold standard evaluation, the semi-latent class approach is more conceptually appealing, more robust under verification completely at random, and allows for model comparisons using likelihoods with only small number of tests.

Acknowledgments

We thank Dr. Iinuma for providing us access to the Gastric Cancer dataset. We thank Dr. Seirchiro Yamamoto for helping us get access to the data as well as for helpful conversations. We thank the Center for Information Technology, NIH, for providing access to the high performance computational capabilities of the Beowulf cluster computer system. We also thank the editor, associate editor, and two reviewers for their thoughtful and constructive comments which have lead to an improved paper.

Figure Legend

Figure 1: Contour plot of relative asymptotic bias in sensitivity and specificity for 50% completely at random verification when the true model is a GRE model with $P_d = 0.20$,

$\sigma_0 = 1.5$, $\sigma_1 = 3$, and $J = 5$. Relative asymptotic bias of sensitivity and specificity is defined as $(SENS^* - SENS)/SENS$ and $(SPEC^* - SPEC)/SPEC$, respectively. The contour plot was generated for sensitivities and specificities over an equally spaced grid ranging from 0.65 to 0.95 with 400 points.

Figure 2: Distribution of Estimates of sensitivity using the FM and GRE model under completely random (CR) verification as well as under over-sampling. Data were simulated under a FM model with $J = 5$, $I = 1000$, $SENS = 0.75$, $SPEC = 0.90$, $P_d = 0.2$, $\eta_1 = 0.5$, and $\eta_0 = 0.2$. 1000 simulated realizations were obtained.

APPENDIX

A. *Expected individual contribution to the Log Likelihood Under a Correct and Misspecified Model*

This is evaluated under the assumption of a common sensitivity and specificity across J tests, where the number of positive tests S is a sufficient statistic. Denote Z_{Sd} as an indicator of whether the individual is verified, has S of J positive tests, and is verified with disease status d. Let X_S be an indicator for an individual not being verified and having S positive tests. Denote T as the true model and M as the assumed model. The expected (under T) log-likelihood of the assumed model M is

$$\begin{aligned} E_T[\log L(\boldsymbol{Y}_i, \boldsymbol{\theta}_M)] &= \sum_{d=0}^{1}\sum_{s=0}^{J} E_T[Z_{sd}]\log[P_M(S=s|D=d)P_M(D=d)] \\ &+ \sum_{s=0}^{J} E_T[X_s]\log[P_M(S=s|D=0)P_M(D=0) \\ &+ P_M(S=s|D=1)P_M(D=1)] + C_v, \qquad (5) \end{aligned}$$

where $P_M(S|D)$ and $P_T(S|D)$ is the conditional distribution for the sum of J binary tests from the assumed and true models, respectively. Additionally, $P_M(D)$ and $P_T(D)$ is the probability of disease under the assumed and true models, respectively, and C_v is a constant corresponding to the verification process. Denote $r_s = P(V_i|S = s)$ as the probability of verification for a particular observed sum s. The expected values $E_T[Z_{sd}]$ and $E_T[X_s]$ can be expressed as

$$E_T[Z_{sd}] = r_s P(S = s|D = d)P(D = d) \tag{6}$$

and

$$E_T[X_s] = (1 - r_s)[P(S = s|D = 1)P(D = 1) + P(S = s|D = 0)P(D = 0)]. \tag{7}$$

B. *Alternative models for the conditional dependence between tests*

Bahadur model: Let π_{ij} be the probability of a positive response conditional on d_i for the jth test on the ith subject. Let $e_{ij} = \frac{Y_{ij}-\pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}}$ and let $\rho_{ijk} = E[e_{ij}e_{ik}|d_i], \rho_{ijkl} = E[e_{ij}e_{ik}e_{il}|d_i], ..., \rho_{ijkl...J} = E[e_{ij}e_{ik}e_{il}...e_{iJ}|d_i]$. The probability distribution can be expressed as $f(\boldsymbol{Y}_i|d_i) = g(\boldsymbol{Y}_i|d_i) \prod_{j=1}^{J} \pi_{ij}^{Y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$, where $g(\boldsymbol{Y}_i) = 1+\sum_{j<k} \rho_{ijk}\, e_{ij}e_{ik}+ \sum_{j<k<l} \rho_{ijkl}e_{ij}e_{ik}e_{il}+ ... + \rho_{ijkl...J}e_{ij}e_{ik}e_{il}...e_{iJ}$.

log-linear model: The probability distribution can be expressed as $f(\boldsymbol{Y}_i|d_i) = \exp\big(\sum_{j=1}^{J}\theta_j Y_{ij}+ \sum_{j<k}\theta_{jk}Y_{ij}Y_{ik} + ... + \theta_{jk...J}Y_{ij}Y_{ik}...Y_{iJ} + \Delta\big)$, where Δ is a normalization factor so that $f(\boldsymbol{y}_i|d_i)$ sum to one over all values of $\boldsymbol{y}_i$.

Beta-binomial model: This distribution assumes that the probability of a positive test (conditional on d_i) is common across the J tests. The probability distribution is $P(\boldsymbol{Y}_i|d_i) = B(S + \alpha, J - S + \beta)/B(\alpha, \beta)$, where $S = \sum_{j=1}^{J} y_{ij}$ which depends on two parameters α and β.

References

Albert, P.S., McShane, L.M., Shih, J.H., et al. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics* **57**, 610-619.

Albert, P.S. and Dodd, L.E. (2004). A cautionary note on the robustness of latent class models for o estimating diagnostic error without a gold standard. *Biometrics* **60**, 427-435.

Alonzo, A. and Pepe, M. (2001). Using a combination of reference tests to assess the accuracy of a diagnostic test. *Statistics in Medicine* **18**, 2987-3003.

Aptech Systems. (1992). *Gauss Systems.* Version 3.0. Kent, Washington; Aptech Systems.

Bahadur, R.R. (1961). A representation of the joint distribution of responses of n dichotomous items. In: *Studies in item analysis and prediction*, H. Solomon (Ed.), Stanford Mathematical Studies in the Social Sciences VI. Stanford, California, Stanford University Press.

Baker, S.G. (1995). Evaluating multiple diagnostic tests with partial verification. *Biometrics* **51**, 330-337.

Begg, C.B. and Greenes, R.A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* **39**, 207-215.

Cox, D.R. (1972). The analysis of multivariate binary data. *Applied Statistics* **21**, 113-120.

Efron, B. and Tibshirani, R.J. (1993). *An introduction to the Bootstrap.* New York: Chapman and Hall.

Heagerty, P.J. and Kurland, B.F. (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika* **88**, 973-985.

Hoenig, J.M., Hanumara, R.C., and Heisey, D.M. (2002). Generalizing double and triple sampling for repeated surveys and partial verification. *Biometrical Journal* **44**, 603-618.

Iinuma, G., Ushiro, K., Ishikawa, T., Nawano, S., Sekiguchi, R., and Satake, M. (2000). Diagnosis of gastric cancer comparison of conventional radiography and digital radiography with a 4 million pixel charge-coupled device. *Radiology* **214**, 497-502.

Irwig, L., Glasziou, P.P., Berry, G., Chock, C., Mock, P., and Simpson J.M. (1994). Efficient study designs to assess the accuracy of screening tests. *American Journal of Epidemiology* **140**, 759-767.

Kosinski, A.S. and Barnhart, H.X. (2003). Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics* **59**, 163-171.

Liang K.Y. and Zeger S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 12-22.

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, Oxford.

Qu Y., Tan, M., and Kutner, M.H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52**, 797-810.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605-610.

Tan, M., Qu, Y., and Rao, J.S. (1999). Robustness of the latent variable model for correlated binary data. *Biometrics* **55**, 258-263.

Tosteson, T.D., Titus-Ernstoff, L., Baron, J.A., and Karagas, M.R. (1994). A two-stage validation study for determining sensitivity and specificity. *Environmental Health Perspectives* **102**, 11-14.

van der Merwe, L. and Maritz, J.S. (2002). Estimating the conditional false-positive rate for semi-latent data. *Epidemiology* **13**, 424-430.

Van Dyck, E., Buve, A., Weiss, H.A. et al. (2004). Performance of commercially available enzyme immunoassays for detection of antibodies against herpes simplex virus type 2 in African populations. *Journal of Clinical Microbiology* **42**, 2961-2965.

Walter, S.D. (1999). Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. *Epidemiology* **10**, 67-72.

Table 1: Estimation of overall prevalence, sensitivity and specificity for digital radiography using no gold standard (GS) and with the consensus rating as the gold standard. Models were fit under the conditional independence (CI) finite mixture (FM), and Gaussian random effects model (GRE) using linuma et al.'s data.

		GS	No GS
P_d	CI	0.24 (0.04)[1]	0.18 (0.04)
	GRE	0.24 (0.04)	0.16 (0.10)
	FM	0.24 (0.04)	0.17 (0.04)
SENS	CI	0.75 (0.06)	0.89 (0.05)
	GRE	0.75 (0.06)	0.92 (0.19)
	FM	0.75 (0.06)	0.91 (0.05)
SPEC	CI	0.91 (0.01)	0.89 (0.02)
	GRE	0.91 (0.01)	0.88 (0.03)
	FM	0.91 (0.01)	0.90 (0.02)

[1] standard errors were estimated using a bootstrap with 1000 bootstrap samples.

Table 2: Estimation of prevalence and rater-specific sensitivity and specificity for digital radiography with no gold standard (GS) and with the consensus rating as the gold standard. Models were fit under the conditional independence (CI), finite mixture (FM), and Gaussian random effects model (GRE)[1] using linuma et al.'s data .

		CI		GRE		FM	
		GS	No GS	GS	No GS	GS	No GS
	P_d Est	0.24	0.18	0.24	0.22	0.24	0.22
	SE	(0.04)[2]	(0.04)	(0.04)	(0.07)	(0.04)	(0.07)
Rater 1	SENS Est	0.67	0.88	0.66	0.77	0.67	0.78
	SE	(0.09)	(0.11)	(0.09)	(0.16)	(0.09)	(0.11)
	SPEC Est	0.99	0.99	0.99	1.00	0.99	1.00
	SE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Rater 2	SENS Est	0.78	0.89	0.77	0.80	0.78	0.81
	SE	(0.08)	(0.07)	(0.08)	(0.14)	(0.08)	(0.08)
	SPEC Est	0.87	0.85	0.87	0.86	0.87	0.86
	SE	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
Rater 3	SENS Est	0.52	0.68	0.51	0.57	0.52	0.57
	SE	(0.10)	(0.13)	(0.10)	(0.14)	(0.10)	(0.13)
	SPEC Est	0.99	0.99	0.99	0.97	0.99	0.99
	SE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Rater 4	SENS Est	0.81	1.00	0.82	0.92	0.81	1.00
	SE	(0.07)	(0.05)	(0.07)	(0.14)	(0.08)	(0.01)
	SPEC Est	0.97	0.95	0.97	1.00	0.97	0.99
	SE	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)
Rater 5	SENS Est	0.85	1.0	0.86	0.92	0.85	0.92
	SE	(0.07)	(0.05)	(0.07)	(0.14)	(0.07)	(0.04)
	SPEC Est	0.72	0.71	0.72	0.72	0.72	0.72
	SE	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)
Rater 6	SENS Est	0.89	0.94	0.88	0.84	0.89	0.85
	SE	(0.06)	(0.07)	(0.07)	(0.09)	(0.06)	(0.06)
	SPEC Est	0.90	0.85	0.89	0.86	0.90	0.86
	SE	(0.03)	(0.04)	(0.03)	(0.05)	(0.03)	(0.04)

[1] There are 13 $(2J+1)$ parameters for the CI model and 15 $(2J+3)$ parameters for the FM and GRE models.

[2] standard errors were estimated using a bootstrap with 1000 bootstrap samples.

Table 3: Large sample robustness of the assumed Gaussian random effects (GRE) model to the true dependence structure between tests given by the finite mixture (FM) model. Verification is independent of $\boldsymbol{Y}_i$, and r denotes the proportion of random samples verified (i.e., $P(V_i = 1) = r$). The true model is a FM with $\eta_0 = 0.2$, $P_d = 0.20$, $SENS = 0.75$, and SPEC=0.9 for differing r and η_1 and with $J = 5$. Asymptotic bias for sensitivity and specificity is $SENS^* - SENS$ and $SPEC^* - SPEC$, respectively.

			Estimators misspecified Model		Expected Log-likelihood[1]	
r	η_1	P_d^*	$SENS^*$	$SPEC^*$	$E_{\text{FM}}[\log L_{\text{FM}}]$	$E_{\text{FM}}[\log L_{\text{GRE}}]$
0	0.2	0.41	0.45	0.92	-2.24106	-2.24106
0.02		0.21	0.71	0.90	-2.24327	-2.24374
0.2		0.20	0.74	0.90	-2.26317	-2.26454
1		0.20	0.75	0.90	-2.35160	-2.35410
0	0.5	0.15	0.77	0.86	-2.10467	-2.10476
0.02		0.16	0.75	0.87	-2.10796	-2.10973
0.20		0.18	0.76	0.89	-2.13749	-2.14648
1		0.20	0.76	0.90	-2.26875	-2.28668

[1] Expected individual contribution to the log-likelihood.

Table 4: Large sample robustness of the assumed finite mixture (FM) model to the true dependence structure between tests given by the Gaussian random effects (GRE) model. Verification is independent of $\boldsymbol{Y}_i$ and r denotes the proportion of random samples verified (i.e., $P(V_i = 1) = r$). The true model is a GRE model with $P_d = 0.2$, $SENS = 0.75$, $SPEC = 0.9$, $\sigma_0 = 1.5$ and $J = 5$ for differing r.

			Estimators misspecified Model		Expected Log-likelihood[1]	
r	σ_1	P_d^*	$SENS^*$	$SPEC^*$	$E_{\text{GRE}}[\log L_{\text{GRE}}]$	$E_{\text{GRE}}[\log L_{\text{FM}}]$
0	1.5	0.22	0.84	0.94	-1.74339	-1.74339
0.2		0.21	0.82	0.93	-1.78383	-1.81920
0.5		0.20	0.78	0.91	-1.86950	-1.91198
1		0.20	0.75	0.90	-1.99560	-2.05440
0	3	0.22	0.86	0.95	-1.61806	-1.61806
0.2		0.21	0.82	0.93	-1.67106	-1.70415
0.5		0.20	0.76	0.90	-1.75057	-1.79748
1		0.20	0.75	0.90	-1.88307	-1.95351

[1] Expected individual contribution to the log-likelihood.

Table 5: Range in relative asymptotic bias for the GRE and FM models when the true conditional dependence structure is (i) Bahadur model with all three and higher way correlations equal to zero, (ii) a log-linear model with a three way interaction, and a (iii) beta-binomial model. The range is over a range of sensitivities and specificities between 0.65 and 0.95. The range in relative bias is for $P_d = 0.20$, $J = 5$, and for 50% completely random verification ($r = 0.5$).

True Model		GRE	FM
Bahadur model[1]	SENS	-2.8% to 0.19%	-0.83% to 4.4%
	SPEC	-0.75% to 0.18%	-0.40% to 1.6%
Log-linear model [2]	SENS	0% to 4.3%	0% to 7.0%
	SPEC	-0.20% to 1.3%	0% to 3.7%
Beta-binomial model [3]	SENS	-0.13% to 0%	0.17% to 4.3%
	SPEC	-0.07% to 0.05%	0.14% to 3.8%

[1] Bahadur model with two-way correlations of 0.20 and all three and higher-way correlations equal to zero.

[2] log-linear model with $logP(\boldsymbol{Y}_i|d_i) = \beta_{d_i} + 0.5I + \Delta$, where I is an indicator which is equal to one if at least three or more of the Y_{ij}'s are equal to one and where Δ is a normalizing constant so that $P(\boldsymbol{Y}_i|d_i)$ sum to one over all possible $\boldsymbol{Y}_i$. The parameters β_{d_i} were chosen to correspond to the different values of sensitivity and specificity.

[3] $P(\boldsymbol{Y}_i|d_i)$ followed beta-binomial distributions with $\beta = 0.4$ (for both $d_i = 0$ or 1) and α varied corresponding to the desired sensitivity or specificity.

Table 6: Large sample robustness of the assumed Gaussian random effects model (GRE) when the true dependence structure between tests is a finite mixture model (FM). Verification is restricted to those patients who screen positive on at least one test and r is the proportion of samples who are verified at random from those patients (i.e., $P(V_i = 1| \sum_{j=1}^{J} y_{ij} = s) = 0$ if $s = 0$ and r, otherwise). The true model is a FM with $\eta_0 = 0.2$, $P_d = 0.2$, SENS=0.75 and SPEC=0.9 for differing r and η_1 and with $J = 5$.

		Estimators misspecified Model			Expected Log-likelihood[1]	
r	η_1	P_d^*	$SENS^*$	$SPEC^*$	$E_{\text{FM}}[\log L_{\text{FM}}]$	$E_{\text{FM}}[\log L_{\text{GRE}}]$
0	0.2	0.41	0.45	0.92	-2.24106	-2.24106
0.02		0.22	0.70	0.90	-2.24359	-2.24319
0.2		0.20	0.73	0.90	-2.26357	-2.26241
1		0.20	0.75	0.90	-2.34997	-2.34782
0	0.5	0.15	0.77	0.86	-2.10467	-2.10476
0.02		0.16	0.74	0.87	-2.10903	-2.10758
0.20		0.25	0.57	0.88	-2.13865	-2.13370
1		0.26	0.57	0.89	-2.25944	-2.24983

[1] Expected individual contribution to the log-likelihood.

Table 7: Simulations with a common sensitivity and specificity. Data was simulated under the finite mixture (FM) model with $P_d = 0.2$, $\eta_0 = \eta_1 = 0.2$, $SENS = 0.75$,$SPEC = 0.90$, and $J = 5$. Results are based on 1000 simulations. Mean parameter estimate and standard errors in ( ) are presented.

		FM model Avg. est.			GRE model Avg est.			
I	r	P_d	$SENS$	$SPEC$	P_d	$SENS$	$SPEC$	$\%(logL_{FM} >> logL_{GRE})$[1]
100	0	0.20	0.75	0.90	0.25	0.64	0.88	2%
		(0.05)	(0.10)	(0.03)	(0.13)	(0.18)	(0.04)	
100	0.05	0.20	0.76	0.90	0.19	0.74	0.88	7%
		(0.05)	(0.09)	(0.02)	(0.09)	(0.13)	(0.04)	
100	0.10	0.20	0.75	0.90	0.19	0.75	0.89	10%
		(0.05)	(0.09)	(0.02)	(0.07)	(0.11)	(0.03)	
100	0.20	0.20	0.75	0.90	0.20	0.75	0.90	12%
		(0.04)	(0.07)	(0.02)	(0.05)	(0.08)	(0.02)	
100	0.50	0.20	0.75	0.90	0.20	0.75	0.90	16%
		(0.04)	(0.06)	(0.02)	(0.04)	(0.06)	(0.02)	
100	1	0.20	0.75	0.90	0.20	0.75	0.90	19%
		(0.04)	(0.05)	(0.02)	(0.04)	(0.05)	(0.02)	
1000	0	0.19	0.75	0.90	0.32	0.57	0.91	12%
		(0.02)	(0.04)	(0.01)	(0.12)	(0.16)	(0.02)	
1000	0.05	0.19	0.75	0.90	0.20	0.73	0.90	45%
		(0.02)	(0.03)	(0.01)	(0.03)	(0.05)	(0.01)	
1000	0.10	0.20	0.75	0.90	0.20	0.74	0.90	55%
		(0.02)	(0.03)	(0.01)	(0.02)	(0.03)	(0.01)	
1000	0.20	0.20	0.75	0.90	0.20	0.74	0.90	64%
		(0.01)	(0.02)	(0.01)	(0.02)	(0.03)	(0.01)	
1000	0.50	0.20	0.75	0.90	0.20	0.75	0.90	75%
		(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	
1000	1	0.20	0.75	0.90	0.20	0.75	0.90	79%
		(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	

[1] Proportion of realizations where the log-likelihood under the FM model is more than 1 larger than the log-likelihood under the misspecified GRE model.

Table 8: Simulations with a common sensitivity and specificity. Data was simulated under the Gaussian random effects (GRE) model with $P_d = 0.2$, $\sigma_0 = \sigma_1 = 1.5$, $SENS = 0.75$, $SPEC = 0.90$, $I = 1000$, $J = 5$. Results are based on 1000 simulations.

	FM model Avg. est.			GRE model Avg est.			
r	P_d	$SENS$	$SPEC$	P_d	$SENS$	$SPEC$	$\%(logL_{GRE} >> logL_{FM})^1$
0	0.24	0.84	0.95	0.20	0.73	0.88	0%
	(0.03)	(0.04)	(0.01)	(0.06)	(0.18)	(0.06)	
0.2	0.22	0.81	0.92	0.20	0.75	0.90	100%
	(0.01)	(0.04)	(0.01)	(0.02)	(0.04)	(0.01)	
0.5	0.21	0.78	0.91	0.20	0.75	0.90	100%
	(0.01)	(0.03)	(0.01)	(0.02)	(0.03)	(0.01)	
0.8	0.21	0.76	0.90	0.20	0.75	0.90	100%
	(0.01)	(0.03)	(0.01)	(0.01)	(0.03)	(0.01)	
1	0.20	0.75	0.90	0.20	0.75	0.90	100%
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	

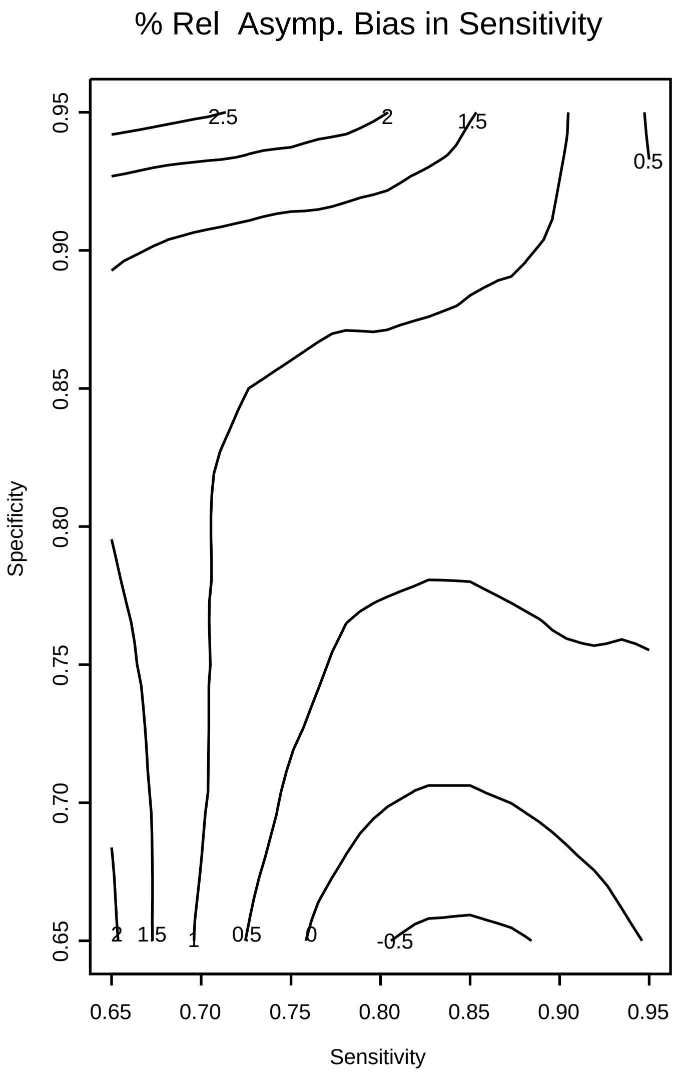
[1] Proportion of realizations where the log-likelihood under the FM model is more than 1 larger than the log-likelihood under the misspecified GRE model.

Table 9: Simulations with four tests with rater-specific sensitivity and specificity. Data was simulated under the finite mixture (FM) model with $P_d = 0.5$, $\eta_0 = \eta_1 = 0.5$, with $SENS = 0.80, 0.85, 0.90,$ and 0.95 and $SPEC = 0.95, 0.90, 0.85,$ and 0.80 for the four tests. Results are based on 1000 simulations.

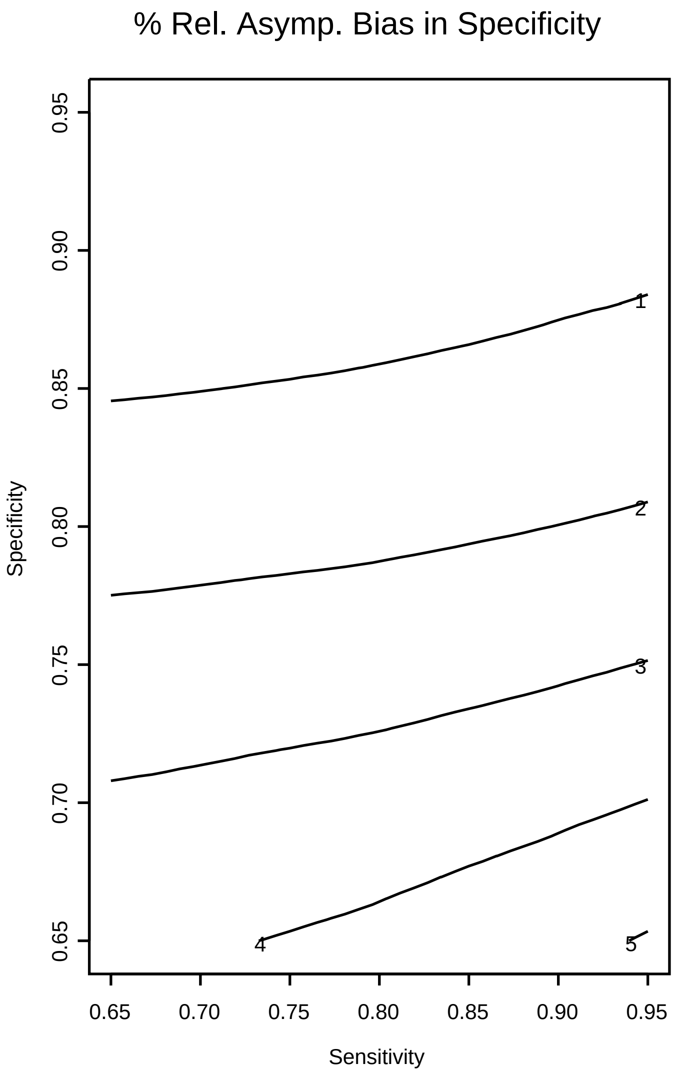
Test		Truth	Avg. Est.					
			$r = 0$		$r = 0.20$		$r = 1$	
			FM	GRE	FM	GRE	FM	GRE
1	SENS	0.80	0.80	0.64	0.80	0.79	0.80	0.80
			(0.08)	(0.23)	(0.03)	(0.03)	(0.02)	(0.02)
	SPEC	0.95	0.95	0.79	0.95	0.94	0.95	0.95
			(0.03)	(0.11)	(0.01)	(0.02)	(0.01)	(0.01)
2	SENS	0.85	0.85	0.72	0.85	0.84	0.85	0.85
			(0.06)	(0.20)	(0.02)	(0.02)	(0.02)	(0.02)
	SPEC	0.90	0.90	0.72	0.90	0.89	0.90	0.90
			(0.05)	(0.10)	(0.02)	(0.02)	(0.01)	(0.01)
3	SENS	0.90	0.90	0.77	0.90	0.89	0.90	0.90
			(0.05)	(0.19)	(0.02)	(0.02)	(0.01)	(0.01)
	SPEC	0.85	0.85	0.68	0.85	0.84	0.85	0.85
			(0.06)	(0.11)	(0.02)	(0.02)	(0.012	(0.02)
4	SENS	0.95	0.95	0.79	0.95	0.94	0.95	0.95
			(0.03)	(0.21)	(0.01)	(0.02)	(0.01)	(0.01)
	SPEC	0.80	0.80	0.64	0.80	0.79	0.80	0.80
			(0.08)	(0.13)	(0.02)	(0.02)	(0.02)	(0.02)
	P_d	0.50	0.50	0.50	0.50	0.50	0.50	0.50
			(0.07)	(0.22)	(0.02)	(0.02)	(0.02)	(0.02)
$\%logL_{\text{FM}} > logL_{\text{GRE}}$			0.63		0.95		0.97	
% Order preserved SENS			0.85	0.74	0.96	0.95	0.98	0.98
% Order preserved SPEC			0.86	0.65	0.96	0.95	0.98	0.98

Table 10: Estimation of overall and rater-specific sensitivity and specificity as well as prevalence for digital radiography using partial verification designs were evaluated. Individuals were re-sampled with replacement to obtain a re-sampled dataset of 113 patients. Verification was done completely at random with probability r. 1000 resampled datasets were obtained and Means (SE) across these datasets are presented. Both the GRE and FM models were used for estimation.

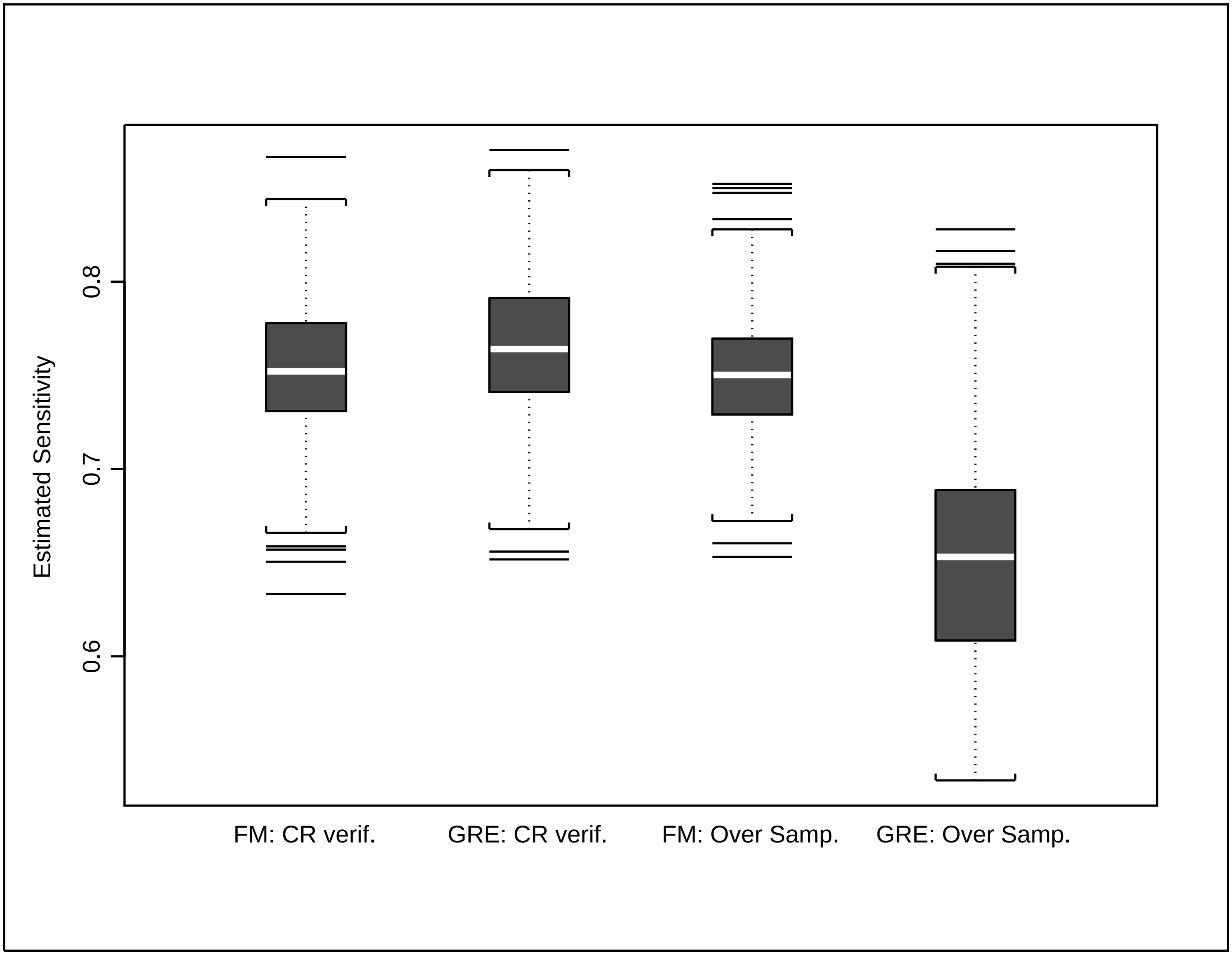
		$r = 0.10$		$r = 0.20$		$r = 0.50$		$r = 0.80$	
Rater		FM	GRE	FM	GRE	FM	GRE	FM	GRE
Overall	P_d	0.21	0.22	0.22	0.23	0.23	0.24	0.24	0.24
		(0.05)	(0.07)	(0.05)	(0.06)	(0.04)	(0.05)	(0.04)	(0.04)
	SENS	0.84	0.80	0.82	0.78	0.78	0.76	0.76	0.76
		(0.08)	(0.14)	(0.08)	(0.11)	(0.07)	(0.07)	(0.06)	(0.06)
	SPEC	0.90	0.89	0.90	0.90	0.91	0.90	0.91	0.91
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
	P_d	0.20	0.23	0.21	0.22	0.23	0.23	0.24	0.24
		(0.05)	(0.06)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)
1	SENS	0.81	0.75	0.76	0.72	0.71	0.69	0.68	0.67
		(0.13)	(0.16)	(0.13)	(0.15)	(0.11)	(0.11)	(0.10)	(0.10)
	SPEC	0.99	1.00	0.99	1.00	0.99	0.99	0.99	0.99
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
2	SENS	0.85	0.76	0.83	0.76	0.80	0.76	0.79	0.77
		(0.09)	(0.13)	(0.09)	(0.12)	(0.09)	(0.10)	(0.08)	(0.09)
	SPEC	0.86	0.86	0.87	0.86	0.87	0.87	0.97	0.87
		(0.04)	(0.05)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)
3	SENS	0.62	0.55	0.57	0.53	0.54	0.52	0.52	0.51
		(0.13)	(0.14)	(0.14)	(0.14)	(0.11)	(0.11)	(0.11)	(0.11)
	SPEC	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
		(0.01)	(0.01)	(0.01)	(0.01)	(0.011	(0.01)	(0.01)	(0.01)
4	SENS	0.96	0.90	0.93	0.88	0.87	0.85	0.83	0.83
		(0.07)	(0.14)	(0.08)	(0.13)	(0.09)	(0.10)	(0.08)	(0.08)
	SPEC	0.97	0.98	0.97	0.98	0.97	0.97	0.97	0.97
		(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
5	SENS	0.95	0.87	0.92	0.86	0.88	0.86	0.86	0.85
		(0.07)	(0.12)	(0.08)	(0.11)	(0.08)	(0.09)	(0.06)	(0.08)
	SPEC	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
		(0.05)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
6	SENS	0.91	0.82	0.90	0.83	0.88	0.85	0.89	0.87
		(0.08)	(0.13)	(0.08)	(0.11)	(0.07)	(0.09)	(0.06)	(0.07)
	SPEC	0.86	0.86	0.87	0.87	0.88	0.88	0.89	0.88
		(0.04)	(0.05)	(0.04)	(0.05)	(0.04)	(0.04)	(0.03)	(0.04)



% Rel Asymp. Bias in Sensitivity



% Rel. Asymp. Bias in Specificity


Estimated Sensitivity
0.6
0.7
0.8
FM: CR verif.
GRE: CR verif.
FM: Over Samp.
GRE: Over Samp.