Imputation approaches for estimating diagnostic accuracy for multiple tests from partially verified designs

(MS # 051230CF)

Paul S. Albert
Biometric Research Branch
Division of Cancer Treatment and Diagnosis
National Cancer Institute
albertp@mail.nih.gov

March 27, 2006

Summary

Interest often focuses on 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. Various authors have proposed semi-latent class modeling approaches for estimating diagnostic error in this situation. This paper presents imputation approaches for this problem. We show how imputation provides a simpler way of performing diagnostic error and prevalence estimation than the use of semi-latent modeling. Furthermore, the imputation approach is more robust to modeling assumptions and, in general, there is only a moderate efficiency loss relative to a correctly specified semi-latent class model. We apply imputation to a study designed to estimate the diagnostic accuracy of digital radiography for gastric cancer. We illustrate the feasibility and robustness of imputation with analysis, asymptotic results, and simulations.

Key words: Diagnostic accuracy; Gold standard evaluation; Mean imputation; Multiple tests; Partial verification; Prevalence; Sensitivity; Specificity, Verification bias.

1 Introduction

Diagnostic tests are often used to diagnose or screen for a disease. Estimates of sensitivity and specificity are commonly used quantities to compare tests as well as to judge the quality of each test individually. Ideally, a series of experimental tests are performed on each patient along with a definite gold standard test (aka, a reference test). Unfortunately, gold standard evaluation may be expensive, time-consuming, or unethical to perform on all subjects. Failure to account for the verification process (possibly resulting in verification bias) is a well known problem which has been extensively studied for the case of single experimental test (Begg and Greenes (1983), Zhou (1993), and an interesting application by Punglia, et al. (2003)). For multiple experimental tests, various authors have proposed semi-latent models which incorporate information on tests with and without gold standard evaluation. Walter (1999) and VanderMerwe and Maritz (2002) focused on the situation of two tests when only individuals with at least one positive test are verified with the gold standard test. Albert and Dodd (2006) proposed two classes of models for the dependence between tests for partially verified designs. Both of these models can easily be applied to either the situation in which individuals are verified completely at random or to the situation in which the probability of verification depends on the actual test results.

Although model-based approaches work well when the model is correctly specified, and in many cases when the model is misspecified, estimation is complex and requires specialized software. Imputation and re-weighting estimators for prevalence, sensitivity, and specificity provide simple alternative to model based approaches, which have been studied for the case of a single experimental test (Gao et al. (2000), Alonzo et al. (2001), Alonzo and Pepe (2005)). There has been little work investigating imputation and re-weighted estimators in the case of multiple correlated tests on a given individual. A notable exception is Zheng et al. (2005) who propose a weighted estimating equations approach in this setting.

This paper discusses imputation approaches for estimating diagnostic error and prevalence with multiple tests under partial verification. The approaches are described in Section 2. In Section 3, we analyze the data from a study evaluating the use of digital radiography for diagnosing gastric cancer using imputation approaches, a re-weighted estimation approach, and alternative semi-latent class modeling approaches. In Section 4, we show the asymptotic bias of the imputation approaches under two semi-latent class models which account for the dependence between tests in different ways. We present the results of simulation studies in Section 5. A discussion follows in Section 6.

2 Imputation Approach

Let $i = 1, 2, ..., I$ be an index for individuals and let $j = 1, 2, ..., J$ be an index for tests. Let d_i be the true disease status, define $P_d = P(d_i = 1)$ as the prevalence of disease, and let v_i be an indicator of whether the ith patient is verified. Further, let y_{ij} be the jth binary experimental test result on the ith patient and $\boldsymbol{Y}_i = (y_{i1}, y_{i2}, ..., y_{iJ})'$. We begin by addressing the problem of estimating a common sensitivity and specificity across J experimental tests. This is appropriate when a single test is repeated multiple times on each individual or when interest focuses on obtaining an average sensitivity and specificity across J tests. Further, the common sensitivity and specificity assumption leads to a simple formulation with a limited number of parameters for examining the statistical properties of the imputation approaches for multiple tests.

When gold standard evaluation is available on all subjects, simple method of moments estimators of prevalence (P_d), sensitivity (Se), and specificity (Sp) are

$$\widehat{P_d} = \frac{1}{I}\sum_{i=1}^{I} d_i, \quad \widehat{Se} = \frac{\sum\limits_{i=1}^{I} y_{i.}d_i}{J\sum\limits_{i=1}^{I} d_i}, \quad \text{and} \quad \widehat{Sp} = \frac{\sum\limits_{i=1}^{I}(J - y_{i.})(1 - d_i)}{J\sum\limits_{i=1}^{I}(1 - d_i)}. \tag{1}$$

These estimators of the common sensitivity and specificity are maximum likelihood estimators under an independence assumption across multiple tests, and are asymptotically unbiased under an arbitrary dependence structure between tests.

We discuss various approaches for imputing values of d_i into (1) when d_i is not observed. These approaches are generalizations of prior methodologies (Guo et al. (2000) and Alonzo et al. (2001) for prevalence and Alonzo and Pepe (2005) for diagnostic accuracy), which are now applied to the situation of multiple experimental tests.

For verified observations ($v_i = 1$), we fit the following simple logistic regression model to $(d_i, \boldsymbol{Y}_i)$ data,

$$\text{logit} P(d_i = 1 | y_{i.}, v_i = 1) = \beta_0 + \beta_1 y_{i.} \ \text{ and } \ y_{i.} = \sum_{j=1}^{J} y_{ij}. \tag{2}$$

The resulting maximum-likelihood estimators are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$, and

$$\widehat{P}(d_i = 1 | y_{i.}, v_i = 1) = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 y_{i.})}{\{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 y_{i.})\}}. \tag{3}$$

For mean imputation (MI), we impute values of $\widehat{P}(d_i = 1|y_{i.}, v_i = 1)$ (dented by $\widehat{d_i}$) in place of the unobserved d_i's in (1) when observations are not verified ($v_i = 0$). For semi-parametric efficient estimation, we replace values of d_i by $\frac{v_i d_i}{(P(v_i=1|\boldsymbol{y}_i)} - (\frac{v_i - P(v_i=1|\boldsymbol{y}_i)}{P(v_i=1|\boldsymbol{y}_i})\widehat{d_i}$, where $P(v_i = 1|\boldsymbol{y}_i)$ is the probability of verification which can depend on the test results $\boldsymbol{y}_i$ (we will discuss the verification mechanism in more detail later). When the verification process is known (we are considering a verification process which is fixed by design), the semi-parametric estimator (SPE) is unbiased under a misspecified model (Alonzo et al. (2001); Alonzo and Pepe (2005)). An alternative to these two imputation approaches is a re-weighted estimator in which d_i is replaced by $v_i d_i / P(v_i = 1|\boldsymbol{y}_i)$. For multiple tests, this re-weighted estimator is closely related to the generalized estimating equations approach proposed by Zeng et al. (2005).

The imputation and re-weighted approaches can easily be adapted for estimating the rater-specific estimates of sensitivity and specificity. Sensitivity and specificity for the jth test (Se_j and Sp_j) can be estimated as

$$\widehat{Se_j} = \frac{\sum\limits_{i=1}^{I} y_{ij} d_i}{\sum\limits_{i=1}^{I} d_i} \quad \text{and} \quad \widehat{Sp_j} = \frac{\sum\limits_{i=1}^{I} (1 - y_{ij})(1 - d_i)}{\sum\limits_{i=1}^{I} (1 - d_i)}. \tag{4}$$

We propose imputing non-verified d_i's into (4) by using a logistic regression model with a regression coefficient corresponding to each of the J test results,

$$\text{logit} P(d_i = 1|y_{i1}, y_{i2}, .., y_{iJ}, v_i = 1) = \beta_0 + \sum_{j=1}^{J} \beta_j Y_{ij}. \tag{5}$$

Specifically, as in the common sensitivity and specificity case, we replace non-observed d_i's with $\widehat{P}(d_i = 1|y_{i1}, y_{i2}, ..., y_{iJ}, v_i = 1)$. More complex logistic regression models with high-order interaction terms corresponding to interactions between multiple tests are possible. However, we found that estimates of diagnostic error and prevalence were nearly unbiased relative to a correctly specified parametric model when using simple additive imputation models. Alternatively, a simpler imputation model such as (2) can be used in place of (5). The advantages and disadvantages of using (2) versus (5) are studies with simulations in Section 5.

Standard errors of $\widehat{P_d}$ as well as a common and rater-specific $\widehat{Se}$, and $\widehat{Sp}$ can be estimated using the Bootstrap (Efron and Tibshirani, 2003). Specifically, we construct a bootstrap sample by sampling I individual observations $(d_i, \boldsymbol{Y}_i, v_i)$ at random with replacement from the I total observations in the original dataset. For each bootstrap sample, estimates of prevalence, sensitivity, and specificity are obtained using imputation. Standard errors can then be estimated as the sample standard deviation of the bootstrap sample estimates.

The imputation approaches will be compared with the two semi-latent models proposed by Albert and Dodd (2006). The first is a Gaussian random effects model in which dependence between repeated tests is introduced through a Gaussian random effect. This model will be referred to as the Gaussian random effects (GRE) model. The second model incorporates dependence between tests through a finite mixture, whereby given an individual's true status, the individual will either be subject to diagnostic error or be always correctly diagnosed. This model will be referred to as the finite mixture (FM) model. Details about these two semi-latent models are presented in the Appendix.

We will examine the performance of the imputation approaches under different types of verification mechanisms. First, we will consider verification which is completely at random. This type of verification occurs if the verification is a simple random sample chosen independently from the test results $\boldsymbol{Y}_i$. For this process, we will denote the proportion of individuals verified as $r = P(v_i = 1)$. Second, we will consider verification in which the probability of verification depends on the test results $\boldsymbol{Y}_i$, where $r_{g(\boldsymbol{Y}_i)} = P\{v_i = 1 | g(\boldsymbol{Y}_i)\}$, and where g is a function of the test results $\boldsymbol{Y}_i$. In this paper, we consider $g(\boldsymbol{Y}_i) = \sum_{j=1}^{J} y_{ij}$. This type of verification has been referred to as verification biased sampling (Begg and Greenes, 1983; Pepe, 2003). An important special case, called extreme verification biased sampling occurs when the reference standard test is obtained only on individuals with at least one positive test because determining the gold standard test requires an invasive procedure such as surgery ($r_0 = 0$ and $r_j = 1$, $j = 1, 2, ..., J$). In this situation, it may not be ethical to obtain gold standard evaluation when all tests are negative.

3 Motivating Example

The methodology is motivated by a medical imaging study which compared conventional versus digital radiography for diagnosing gastric cancer (Iinuma, et al., 2000). In this study, six radiologists ($J = 6$) evaluated 225 images on either conventional ($I = 112$) or digital ($I = 113$) radiography in order to compare 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. The review was done on 225 patients, although such an extensive consensus review may not be possible for larger studies. We use the digital radiography data from this study to illustrate the imputation approach and to compare the results with those obtained by fitting the GRE and the FM models presented in Albert and Dodd (2006) (see Appendix).

Table 1 shows the estimated diagnostic error for digital radiography using mean imputation for a differing proportions of completely random verification (r) and for extreme biased sampling ($r_0 = 0$, $r_s = 1$, $s = 1, 2, 3, ..., 6$). The table shows estimated prevalence, sensitivity, and specificity for the common sensitivity and specificity as well as for rater-specific sensitivities and specificities. For the common sensitivity and specificity we imputed data based on (3) and for rater-specific estimates of sensitivity and specificity we imputed non-observed d_i's based on (4) (Results were very similar when we imputed based on (3) when estimating rater-specific sensitivity and specificity. We will examine the effect of using a simpler imputation model versus a more complex one in Section 5). In order to capture the variability associated with different amounts of verification, we resampled data with replacement and incorporated the reference standard on a given image subject to the verification mechanism. Means and standard errors summarizing the 1000 bootstrap samples are presented in Table 1. The results show that the means are close across the different verification mechanisms. Specifically, means computed under completely random verification with $r = 0.20$ and $r = 0.50$ as well as extreme biased sampling are close to estimates obtained under complete verification ($r = 1$). For extreme biased verification, it was particularly surprising that mean imputation performed so well, since an extrapolation of the logistic imputation model to the case of all negative tests is required. We will show, in the next section, that this extrapolation works well under certain probability mechanisms, but poorly under others.

Under complete verification, the mean imputation approach reduces to a method-of-moments estimation approach similar to generalized estimating equations (Liang and Zeger, 1986) under an independence correlation structure assumption. In this case, estimates and standard errors were nearly identical to those obtained with the two semi-latent models under complete verification (Albert and Dodd, 2006). This was true for both the common diagnostic error model and the rater-specific diagnostic error model. With 20% completely random verification, estimates using the imputation approach appear to be closer to the results for complete verification than when we used either the GRE or FM models. For the common diagnostic error model, estimates of P_d, Se, and Sp were 0.23 (SE=0.06), 0.78 (0.07), and 0.91 (0.02) for the GRE model and 0.22 (0.05), 0.82 (0.08), and 0.90 (0.01) for the FM model, respectively.

We estimating diagnostic accuracy and prevalence under both re-weighted estimation and semi-parametric efficient estimation. For completely at random verification, the re-weighted estimation is equivalent to discarding non-verified cases. For the common sensitivity and specificity case under 50% completely random verification, estimates of prevalence, sensitivity, and specificity were 0.24 (SE=0.063), 0.75 (0.083), and 0.91 (0.020) for the re-weighted approach and 0.24 (SE=0.046), 0.75 (0.074), and 0.91 (0.016) for semi-parametric efficient estimation. These results suggest that the weighted estimation is substantially less efficient (50% less efficient) and the semi-parametric efficient estimation is moderately less efficient (approximately 27% less efficient) compared with mean imputation (presented in Table 1). Similar results were observed when we estimated rater-specific sensitivity and specificity (data not shown).

A comparison of the standard errors across the different proportions of verification in Table 1 suggest that, particularly for estimating prevalence and specificity, there is little gain in efficiency in verifying more than 50% of cases. If obtaining reference standard tests is the constraining aspect of a study, these results suggest that it may be more cost-effective to invest reference samples on new patients as compared with observing a larger proportion of verified samples on existing patients.

We also examined other verification biased mechanisms in addition to extreme verification bias. Specifically, we examined a verification mechanism in which we oversampled discrepant cases by verifying images when there was a discrepancy among any of the 6 raters with probability 0.40 and where we verified cases without any discrepancies with probability 0.05. Under this verification mechanism, prevalence, sensitivity, and prevalence are estimated as 0.24 (SE=0.045), 0.76 (0.074), and 0.91 (0.018) for mean imputation and 0.24 (SE=0.65), 0.76 (0.117), and 0.90 (0.021) for semi-parametric efficient estimation. Thus, there is potentially sizable efficiency loss for the semi-parametric efficient estimation relative to mean imputation. Similar results were found for estimation of rater-specific sensitivity and specificity (data not shown).

In all our analyses, we found that mean imputation and semi-parametric efficient estimation led to similar estimates (and both were similar to estimates obtained with the semi-latent class models). The differences in the approaches were in the estimated standard errors. Model based approaches were most efficient (smallest standard errors) followed by mean imputation, semi-parametric efficient estimation, and finally by re-weighted estimation. As we will further demonstrate with asymptotic results and simulations, the imputation approaches provide a good trade-off between (i) simplicity and robustness and (ii) efficiency as compared with semi-latent class modeling approaches. In the next section we examine the asymptotic bias for the mean imputation estimators for the case of a common sensitivity and specificity, under both a completely random and a verification biased sampling mechanism.

4 Asymptotic Bias

Asymptotic bias of the mean imputation estimator were evaluated under both GRE and FM models (note that the semi-parametric efficient estimator is unbiased under any model). Denote $v_i(y_{i.})$ as an indicator that the ith subject is verified which may depend on the sum of the J tests. As the number of individuals I becomes large, the estimator of prevalence, $\widehat{P_d}$ converges to P_d^*, where

$$\begin{aligned} P_d^* &= E\big[d_i v_i(y_{i.}) + \widehat{P}(d_i = 1|y_{i.}, v_i = 1)\{1 - v_i(y)\}\big] \\ &= P_d E\big[P(v_i = 1|y_{i.})\big] + E\big[P^*(d_i = 1|v_i = 1, y_{i.})\{1 - P(v_i = 1|y_{i.})\}\big], \qquad (6) \end{aligned}$$

where $P^*(d_i = 1|v_i = 1, y) = \exp(\beta_0^* + \beta_1^* y)/\{1 + \exp(\beta_0^* + \beta_1^* y)\}$, $\boldsymbol{\beta^*} = (\beta_0, \beta_1)'$, and where

$$\begin{aligned} \boldsymbol{\beta^*} &= arg\max_{\boldsymbol{\beta^*}} E\big\{LogL(\boldsymbol{\beta})\big\} \\ &= E\big[P(v_i = 1|y_{i.})P(d_i = 1|y_{i.}, v_i = 1) log P^*(d_i = 1|v_i = 1, y_{i.}, \boldsymbol{\beta}) \\ &+ P(v_i = 1|y_{i.})\{1 - P(d_i = 1|y_{i.}, v_i = 1\} \\ &x \;\; log\{1 - P^*(d_i = 1|v_i = 1, y_{i.}, \boldsymbol{\beta})\}\big]. \qquad (7) \end{aligned}$$

The expectations in the above equations are taken with respect to the true GRE or FM model.

The estimators of sensitivity and specificity, $\widehat{Se}$ and $\widehat{Sp}$, converge to Se^* and Sp^*, where

$$Se^* = \frac{1}{JP_d^*} E\big[yP(v_i = 1|y_{i.})P(d_i = 1|y_{i.}, v_i = 1) + yP^*(d_i = 1|v_i = 1, y_{i.}, \boldsymbol{\beta}^*)\{1 - P(v_i = 1|y_{i.})\}\big] \qquad (8)$$

and

$$Sp^* = \frac{1}{J(1 - P_d^*)} E\big[(J - y)P(v_i = 1|y_{i.})\{1 - P(d_i = 1|y_{i.}, v_i = 1)\}$$

$$+ \quad (J-y)\{1-P^*(d_i=1|v_i=1,y_{i.},\boldsymbol{\beta}^*)\}\{1-P(v_i=1|y_{i.})\}\big]. \tag{9}$$

Using (5)-(8), we computed the asymptotic relative bias for $\widehat{P_d}$, $\widehat{Se}$, and $\widehat{Sp}$ (defined as $\{P_d^*-P_d\}/P_d$, $\{Se^*-Se\}/Se$, and $\{Sp^*-Sp\}/Sp$, respectively) under various situations. Figure 1 shows the asymptotic relative bias for $\widehat{P_d}$, as a function of the proportion verified and the number of tests J, under the GRE model with completely random verification. The figure shows that, in this case, the asymptotic bias is negligible for any number of tests J. In general, we found that the asymptotic bias for $\widehat{P_d}$, $\widehat{Se}$, and $\widehat{Sp}$ under either a GRE or FM model was negligible with completely random verification.

We found no asymptotic bias under a FM model under verification biased sampling (including extreme biased sampling). However, there is substantial bias under a GRE model. Figure 2 shows the asymptotic relative bias under the GRE model with a random verification process in which $r_0=0.05$, $r_x=0.40$, for $x=1,2,..,J-1$, and $r_J=0.05$. This particular mechanism is one in which discrepant measurements are over-sampled relative to concordant measurements. For this configuration, the prevalence is positively biased, while the sensitivity is negatively biased. In addition, the bias increases with an increasing J. We investigated the asymptotic bias of the imputation procedure under extreme biased sampling under a GRE model with $P_d=0.20$, $Se=0.75$, $Sp=0.90$, and $\sigma_0=\sigma_1=2$. We found sizable asymptotic bias in this situation (we calculated $P_d^*=0.22$, $Se^*=0.69$, and $Sp^*=0.89$). Thus, there is the potential for sizable amounts of bias with mean imputation under extreme biased sampling.

We investigated a slight alteration to the imputation procedure under a verification biased sampling mechanism. Rather than fitting an unweighted logistic regression for imputation, we investigated the performance of a weighted logistic regression in which we weight inversely proportion to the probability of verification. Specifically, we weight each verified observation by $1/P(v_i=1|y_{i.})$. Figure 3 shows that the asymptotic relative bias is negligible when we used the weighted logistic regression predictor. Further evaluations demonstrated that, with the exception of extreme biased sampling, the use of the weighted predictor did very well for all random verification mechanisms. Inverse proportional weighting does not work for extreme biased sampling since, there is no gold standard evaluation when $\sum_{j=1}^{J} y_{ij} = 0$, and therefore, there is no appropriate weight for a category that is never observed.

5 Simulations

A series of simulations were performed to show the finite sample properties of the imputation approaches and re-weighted estimation relative to both a correctly specified semi-latent model and a misspecified one.

Table 2 shows the performance of the imputation approaches relative to both FM and GRE models for a common sensitivity and specificity when, data are simulated under a GRE model, the logistic regression model (2) is used for imputation, and when $I = 1000$ and $J = 5$. Further, data were generated under a verification biased sampling mechanism in which discordant test results were sampled more frequently than concordant results ($r_0 = r_5 = 0.05$ and $r_s = 0.40$ for $s = 1, 2, 3, \text{and } 4$). As reported in Albert and Dodd (2006), the results suggest that, under a verification biased sampling mechanism, estimates of prevalence and diagnostic error obtained using a semi-latent model may be sensitive to the assumed dependence structure between tests. The imputation estimates were computed using the weighted logistic regression where the weights were chosen inversely proportional to the verification probabilities. Recall, that the simple logistic imputation model resulted in biased (asymptotically) estimation when we did not implement this weighting. The results suggest that the imputation estimates are nearly unbiased and that there is only a moderate efficiency loss relative to using the correctly specified semi-latent model (the sensitivity, specificity, and prevalence with the mean imputation estimators are $(0.051/0.066)^2 = 60\%, 60\%$, and 71% as efficient as using the correctly specified GRE model). The re-weighted estimation approach is very inefficient relative to the imputation approaches or model-based approaches. For example, the re-weighted estimators of sensitivity, specificity, and prevalence are $(0.051/.092)^2 = 31\%, 31\%$, and 17% as efficient as using the correctly specified GRE model. Further, although the semi-parametric estimator is robust under a misspecified imputation model, it is moderately less efficient than mean imputation (the efficiency of SPE relative to MI is 89%, 81%, and 86% for sensitivity, specificity, and prevalence, respectively).

We examined the effect of using a seriously misspecified imputation model on estimation. To do this, we assumed the logistic model (2) in which $y_{i.}^2$ replaced $y_{i.}$ and simulated 1000 datasets under a GRE model. (Recall that results in Section 3 suggested that under a GRE model, mean imputation is nearly unbiased with an imputation model given by (2) and when the imputation is done by weighting each observation inversely proportional to the probability of verification.) Similar to simulations presented in Tables 3 and 4, sensitivity, specificity, and prevalence were chosen as 0.75, 0.90, and 0.20. Average MI estimates of sensitivity, specificity, and prevalence were 0.72 (SE=0.066), 0.90 (0.019), and 0.21 (0.026), demonstrating slight bias for sensitivity under the poor imputation model. As expected, average SPE estimates were nearly unbiased with only a slight decrease in efficiency as compared to those presented for the SPE estimates under imputation model (2) in Table 3 (data not shown).

Table 3 shows the performance of the imputation approach under a FM model with the same verification biased sampling mechanism as specified for Table 2. The results are similar to those discussed for Table 2. Specifically, estimates of the semi-latent model under a misspecified model are biased, and the imputation approach provides nearly unbiased estimates with only small efficiency loss relative to using the correctly specified modeling approach. The efficiency of the mean imputation estimator of sensitivity, specificity, and prevalence relative to the correctly specified semi-latent class model is 78%, 97%, and 89%, respectively.

Table 4 shows the performance of the imputation approach for rater-specific estimates of diagnostic error with $I = 1000$ and $J = 4$ under verification biased sampling with a GRE model. Data were imputed using a weighted logistic regression model (weights are inversely proportional to verification probabilities) with linear terms corresponding to the four test results (i.e., using (5)). As in the common sensitivity and specificity case, the MI approach is nearly unbiased and moderately efficient as compared with the true semi-latent GRE model. Note that the misspecified FM model results in substantial bias demonstrating the lack of robustness to modeling assumptions for the semi-latent class models under verification biased sampling. Table 4 also suggests that there is a small loss of efficiency in using the SPE (average relative efficiency of SPE to MI was 93%) and a substantially larger one when using RW estimation, both relative to MI. We also conducted another simulation similar to that presented in Table 4, in which we replaced the more complex linear predictor (5) by the simpler logistic model (2). The results for the MI estimator suggests moderate efficiency in estimating diagnostic accuracy and prevalence at the expense of a slight increase in bias (data not shown). The root-MSE averaged over all four rater-specific sensitivities and specificities as well as over prevalence was slightly lower for the simpler model (avg root MSE was 0.0363 using (2) and 0.0771 using (5)), suggesting that there may not be substantial gains in using the more complex imputation model. Further, there were few differences in the bias and variances of SPE estimators between the two imputation models (data not shown).

6 Discussion

In this paper, we present imputation approaches for estimating diagnostic error and prevalence when there are multiple experimental tests and only a fraction of the individuals have gold standard evaluation. The imputation approaches are shown to have little or no bias and are moderate to highly efficient as compared to a correctly specified semi-latent class model.

There are a number of advantages of imputation over semi-latent class modeling approaches. First, imputation only involves fitting logistic regression models, making implementation simple. The semi-latent class modeling approaches require specialized software, making their implementation more difficult. Second, imputation appears to be more robust than misspecified semi-latent class models. This is particularly important under verification biased sampling where model-based inferences on diagnostic error and prevalence may be sensitive to modeling assumptions (see Tables 2, 3, and 4).

One disadvantage of imputation is under extreme biased sampling, in which individuals who are negative on all tests have no gold standard evaluation. In this case, the imputation approach requires an extrapolation of the logistic regression imputation model to the case where all the experimental test results are negative. Although, this extrapolation appeared to work well in our example (estimates of diagnostic accuracy for extreme-biased sampling were close to those obtained with complete verification), it may not work well under certain probability models. The asymptotic results showed that mean imputation resulted in nearly unbiased estimation for FM models, but could be highly biased under GRE models.

Another potential disadvantage of imputation is the moderate efficiency loss relative to a correctly specified semi-latent class model. In some situations, the extra effort in fitting a semi-latent class model and performing model diagnostics to help assure that the dependence structure between tests is correctly specified may be warranted. However, the additional efficiency of the semi-latent class models may come at the expense of bias if in fact the model is misspecified.

Our analyses and simulation studies suggested that there may be little pay-off in verifying more than 50% of cases. Although various authors have considered optimal design strategies for the case of a single diagnostic test (Irwig, et al. (1994); Tosteson et al. (1994); McNamee (2002)), there has been little or no work on the optimal design problem for multiple tests. This is an area for future research. However, the choice of an optimal design will depend heavily on assumed probability distributions for the multiple tests. Therefore, we question the practicality of developing an optimal design in this situation.

We compared two classes of imputation estimators (mean imputation and semi-parametric efficient estimation) with re-weighted estimation as well the fully efficient semi-latent class modeling approaches. We found the re-weighting approach particularly inefficient for our problem. Although, this loss in efficiency has been reported previously for the case of a single diagnostic test (Alonzo, et al., 2003), there appears to be a larger loss for case of multiple experimental tests. Presumably, this is because (i) the re-weighted approach does not explictly use information about the multiple experimental tests on non-verified cases, and (ii) there is more information in these multiple tests than in a single experimental test. Alonzo et al. (2003) reported efficiency gains for the re-weighted approach when known verification weights were estimated from the data. We would expect similar small gains in efficiency for multiple tests when estimating these verification weights. However, even with these anticipated gains, we would not recommend re-weighted estimation in this setting due to overall low efficiency compared with the other methods considered. We showed that mean imputation resulted in nearly unbiased estimation under two different broad classes of semi-latent class models. Further, although semi-parametric efficient estimation is unbiased in this setting (i.e., unbiased when the verification mechanism is known), there can be moderate efficiency loss relative to estimation with mean imputation.

Our major focus was on the situation where the verification process was fixed by design (e.g., limited resources require the investigators to verify only a small fraction of cases). However, in many situations, investigators attempt to verify everyone, but for various reasons, many individual refuse the gold standard test. Incorporating subject-specific covariates into the imputation model may be appropriate in these cases.

Acknowledgments

We thank Dr. Iinuma and Dr. Seirchiro Yamamoto for providing us access to the Gastric cancer dataset. I also thank Dr. Lori Dodd for helpful discussions on this topic. This research utilized the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, MD. (http://biowulf.nih.gov). I thank the Associate Editor and Reviewer for constructive comments which significantly improved this manuscript.

Appendix: Gaussian random effects (GRE) and finite mixture (FM) models.

As described in Albert and Dodd (2006), the semi-latent models are based on the likelihood

$$L = \prod_{i=1}^{I} \Big[P(\boldsymbol{Y}_i|d_i)P(d_i) \Big]^{v_i} \Big[\sum_{l=0}^{1} P(Y_i|d_i = l)P(d_i = l), \Big]^{1-v_i} \tag{10}$$

where the $P(\boldsymbol{Y}_i|d_i)$ is parameterized to incorporate conditional dependence between tests.

The GRE model incorporates dependence between tests by assuming that $(Y_{ij}|d_i, b_i)$ are independent Bernoulli with proportions given by $\Phi(\beta_{jd_i} + \sigma_{d_i} b_i)$, where the random variable b_i is standard normal and Φ is the cumulative distribution function of a standard normal distribution. Under this model $P(\boldsymbol{Y}_i|d_i) = \int \{\prod_{j=1}^{J} P(Y_{ij}|d_i, b_i)\}\phi(b)db$, 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.

The FM model incorporates dependence between tests through a finite mixture component. Let l_{id_i} be an indicator of whether the ith subject, given disease status d_i, is always classified correctly. Further, let $\eta_0 = P(l_{i0} = 1)$ and $\eta_1 = P(l_{i1} = 1)$. Test results Y_{ij} given d_i and l_{id_i} are assumed independent Bernoulli with probability

$$P(Y_{ij} = 1|d_i, l_{id_i}) = \begin{cases} 1 & \text{if } d_i = 1 \text{ and } l_{i1} = 1 \\ 0 & \text{if } d_i = 0 \text{ and } l_{i0} = 1 \\ \omega_j(1) & \text{if } d_i = 1 \text{ and } l_{i1} = 0 \\ 1 - \omega_j(0) & \text{if } d_i = 0 \text{ and } l_{i0} = 0, \end{cases} \tag{11}$$

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 $l_{i0} = 0$) and the true disease state is d_i. 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. Estimation is based on maximizing (9).

Figure Legends:

Figure 1: Asymptotic relative bias of mean imputation for estimating P_d under a Gaussian random effects (GRE) model with completely random verification. The Gaussian random effects model has $SENS = 0.75$, $SPEC = 0.90$, $P_d = 0.20$, and $\sigma_0 = \sigma_1 = 2$.

Figure 2: Asymptotic relative bias of mean imputation for estimating P_d (solid line), Se (dotted line), and Sp (dashed-dotted line) under a GRE model, verification biased sampling, and an unweighted logistic imputation model. The Gaussian random effects model has

$SENS = 0.75$, $SPEC = 0.90$, $P_d = 0.20$, and $\sigma_0 = \sigma_1 = 2$. The verification mechanism is $r_0 = r_J = 0.05$, $r_x = 0.40$, $x = 1, 2, ..., J-1$.

Figure 3: Asymptotic relative bias for estimating P_d (solid line), Se (dotted line), and Sp (dashed-dotted line) with mean imputation under a GRE model, verification biased sampling, and a weighted logistic imputation model. The weights are chosen inversely proportional to the verification probabilities. The Gaussian random effects model has $SENS = 0.75$, $SPEC = 0.90$, $P_d = 0.20$, and $\sigma_0 = \sigma_1 = 2$. The verification mechanism is $r_0 = r_J = 0.05$, $r_x = 0.40$, $x = 1, 2, ..., J-1$.

References

Albert, P.S. and Dodd, L.E. (2006). On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. In revision at *The Journal of the American Statistical Association.*

Alonzo, T.A., Pepe, M.S., and Lumley, T.S. (2001). Estimating disease prevalence in two-phase studies. *Biostatistics* **4**, 313-326.

Alonzo, T.A. and Pepe, M.S. (2005). Assessing accuracy of a continuous screening test in the presence of verification bias. *Applied Statistics* **54**, 173-190.

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

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

Gao, S., Hui, S.L., Hall, K.S., and Hendrie, H.C. (2000). Estimating disease prevalence from two-phase surveys with non-response at the second phase. *Statistics in Medicine* **19**, 2101-2114.

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., Gasziou, P.P., Berry, G., Chock, C., Mock, P., Simpson, J.M. (1994). Efficient study designs to assess the accuracy of screening tests. *American Journal of Epidemiology* **140**, 759-769.

McNamee, R. (2002). Optimal designs of two-stage studies for estimation of sensitivity and specificity and positive predictive value. *Statistics in Medicine* **21**, 3609-3625.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika.* **73**, 13-22.

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

Punglia, R.S., D'Amico, A.V., Catalona, W.J., Roehl, K.A., and Kuntz, K.M. (2003). Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. *The New England Journal of Medicine* **349**, 335-341.

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

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

Zheng, Y., Barlow, W.E., and Cutter, G. (2005). Assessing accuracy of mammography in the presence of verification bias and intrareader correlation. *Biometrics* **61**, 259-268.

Zhou, X.H. (1993). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communs Statist. Theory Meth*, 22, 3177-3198.

Table 1: Mean imputation estimation of overall and rater-specific sensitivity and specificity as well as prevalence for digital radiography using partial verification designs. Individuals were re-sampled with replacement to obtain a re-sampled dataset of 113 patients. For the first three columns, verification was done completely at random with probability r. For the fourth column, verification was done on all subjects with at least one positive test (extreme biased sampling). 1000 resampled datasets were obtained and means (SE) across these datasets are presented.

Rater		$r = 0.20$	$r = 0.50$	$r = 1.0$	Extrem. bias samp*
Overall	P_d	0.24	0.24	0.24	0.24
		(0.05)	(0.04)	(0.04)	(0.04)
	SENS	0.76	0.76	0.75	0.77
		(0.08)	(0.06)	(0.06)	(0.06)
	SPEC	0.91	0.91	0.91	0.91
		(0.02)	(0.01)	(0.01)	(0.01)
	P_d	0.24	0.24	0.24	0.24
		(0.07)	(0.05)	(0.04)	(0.05)
1	SENS	0.68	0.67	0.67	0.67
		(0.17)	(0.12)	(0.09)	(0.10)
	SPEC	0.99	0.99	0.99	0.99
		(0.02)	(0.01)	(0.01)	(0.01)
2	SENS	0.78	0.77	0.78	0.79
		(0.15)	(0.11)	(0.08)	(0.09)
	SPEC	0.87	0.87	0.87	0.87
		(0.05)	(0.04)	(0.04)	(0.04)
3	SENS	0.52	0.52	0.52	0.52
		(0.15)	(0.11)	(0.10)	(0.10)
	SPEC	0.99	0.99	0.99	0.99
		(0.01)	(0.01)	(0.01)	(0.01)
4	SENS	0.82	0.81	0.82	0.83
		(0.17)	(0.11)	(0.08)	(0.09)
	SPEC	0.97	0.97	0.97	0.97
		(0.03)	(0.02)	(0.02)	(0.02)
5	SENS	0.85	0.85	0.85	0.86
		(0.14)	(0.09)	(0.07)	(0.08)
	SPEC	0.72	0.72	0.72	0.72
		(0.07)	(0.05)	(0.05)	(0.05)
6	SENS	0.85	0.87	0.89	0.90
		(0.14)	(0.09)	(0.06)	(0.08)
	SPEC	0.89	0.89	0.90	0.90
		(0.06)	(0.04)	(0.03)	(0.03)

* The average proportion of individuals verified (at least one positive test result) was 55%.

Table 2: Simulation: A comparison of mean imputation (MI), semi-parametric efficient (SPE), and re-weighted (RW) estimation approaches with a correctly specified GRE model and a misspecified FM model. Equation (2) was used to impute data. Data are generated under a GRE model where $P_d = 0.2$, $\sigma_0 = \sigma_1 = 1.5$, $SENS = 0.75$, $SPEC = 0.90$, $I = 1000$, and $J = 5$. Further, the verification is at random where $r_0 = r_5 = 0.05$ and $r_x = 0.40$ for $x = 1, 2, 3, \text{and } 4$. 1000 resampled datasets were obtained and means (SE) across these datasets are presented.

Method	$SENS$	$SPEC$	P_d
MI	0.75	0.90	0.20
	(0.066)	(0.018)	(0.025)
SPE	0.75	0.90	0.20
	(0.070)	(0.020)	(0.027)
RW	0.74	0.90	0.20
	(0.092)	(0.025)	(0.051)
GRE	0.74	0.90	0.20
	(0.051)	(0.014)	(0.021)
FM	0.83	0.92	0.21
	(0.023)	(0.008)	(0.014)

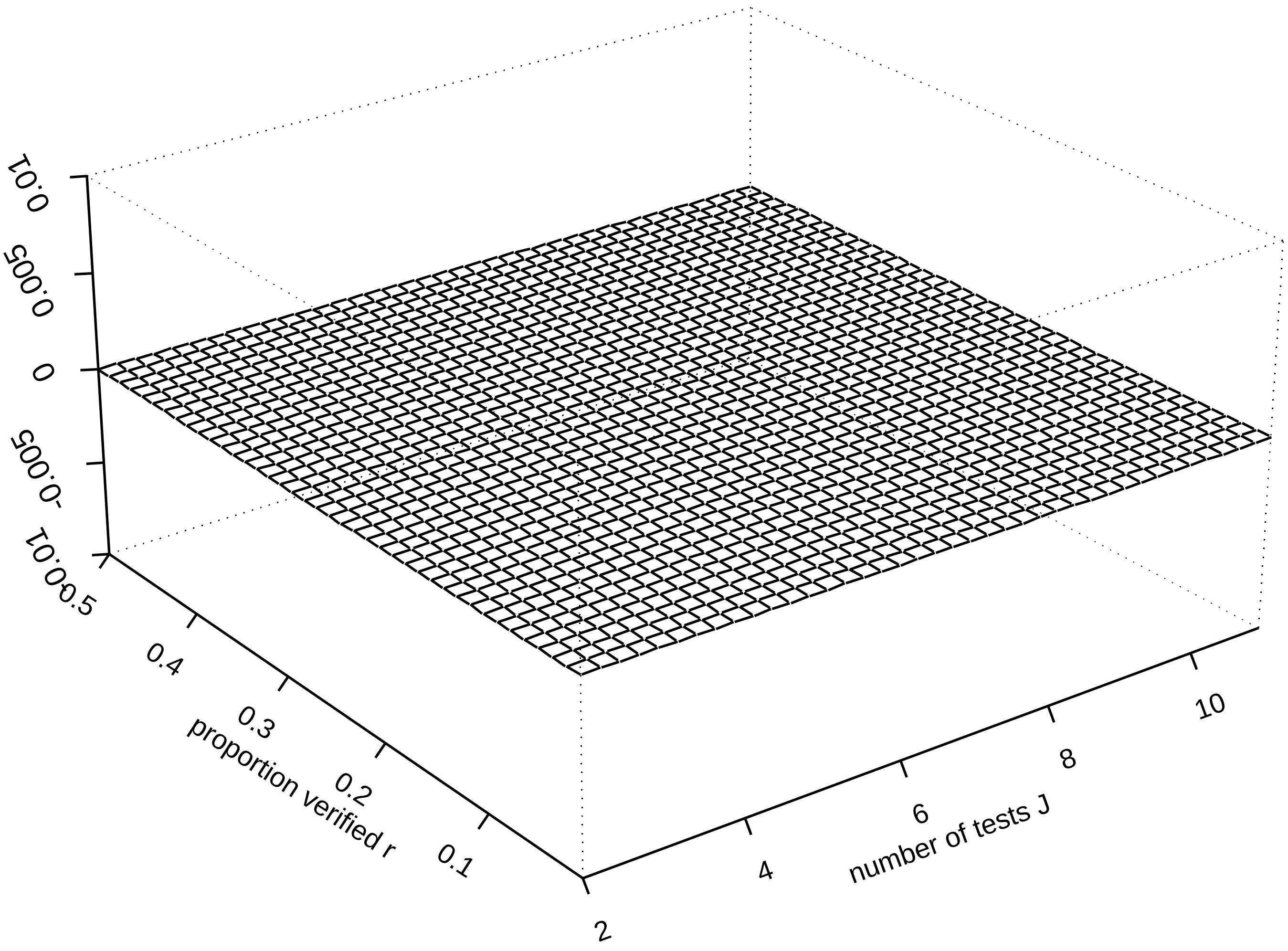
Table 3: Simulation: A comparison of the mean imputation (MI), semi-parametric efficient (SPE), and re-weighted (RW) estimation approaches with a correctly specified FM model and a misspecified GRE model. Equation (2) was used to impute data. Data are generated under a FM model where $P_d = 0.2$, $\eta_0 = 0.2$, $\eta_1 = 0.50$, $SENS = 0.75$, $SPEC = 0.90$, $I = 1000$, and $J = 5$. Further, the verification is at random where $r_0 = r_5 = 0.05$ and $r_x = 0.40$ for $x = 1, 2, 3,$ and 4. 1000 resampled datasets were obtained and means (SE) across these datasets are presented.

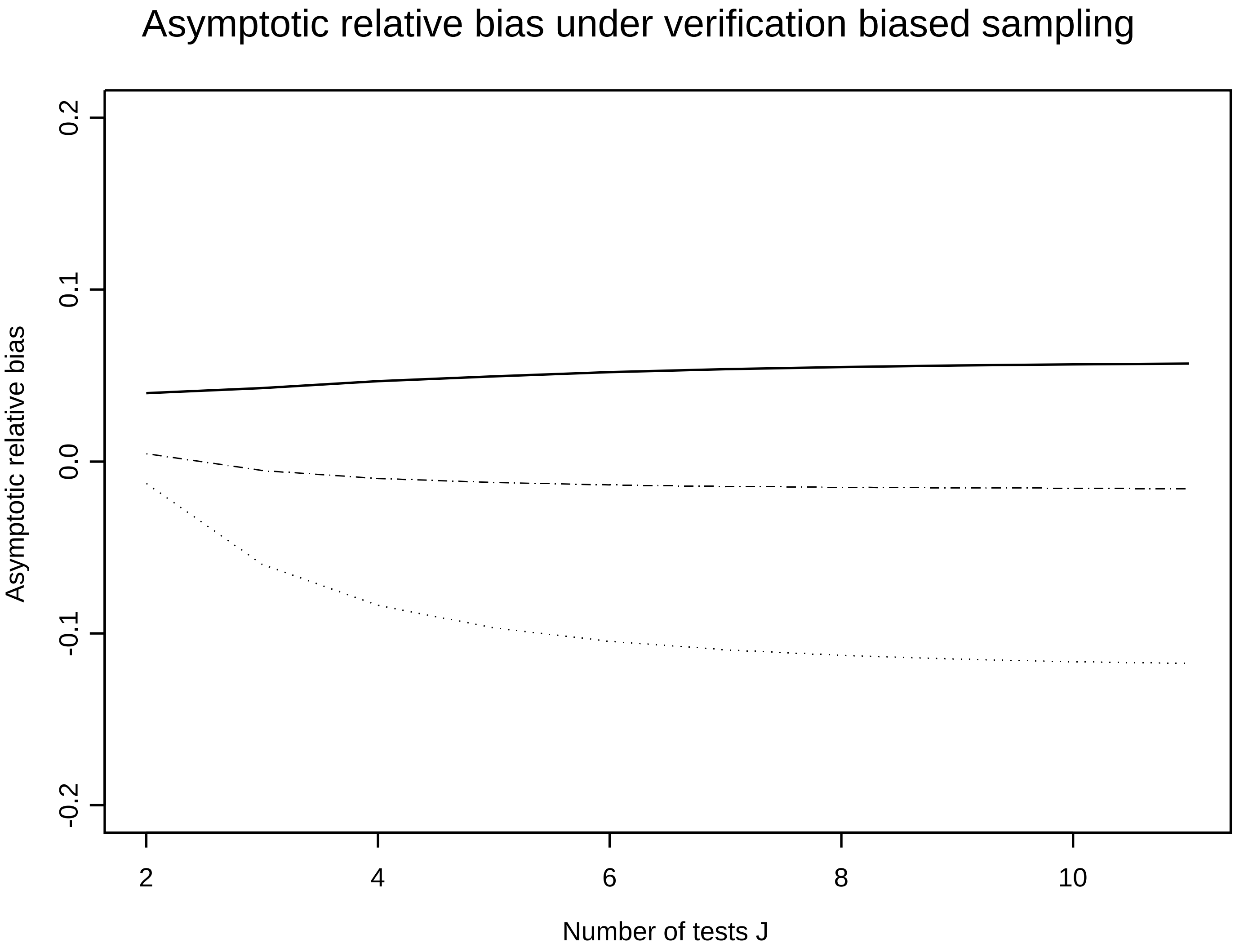
Method	$SENS$	$SPEC$	P_d
MI	0.75	0.90	0.20
	(0.037)	(0.0062)	(0.016)
SPE	0.75	0.90	0.20
	(0.040)	(0.061)	(0.017)
RW	0.74	0.90	0.20
	(0.073)	(0.014)	(0.048)
FM	0.75	0.90	0.20
	(0.030)	(0.0056)	(0.016)
GRE	0.65	0.89	0.23
	(0.051)	(0.0067)	(0.023)

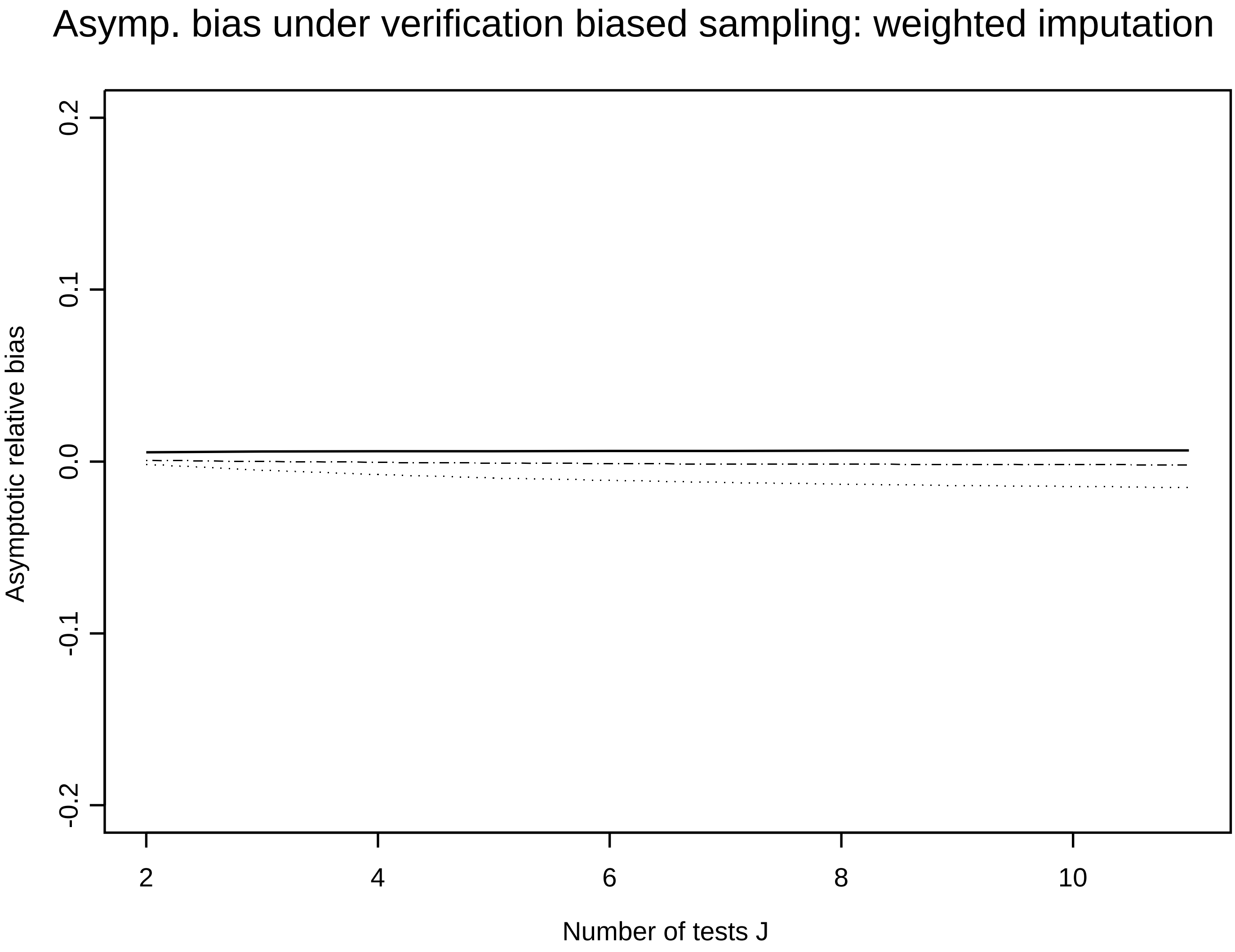
Table 4: Simulation: A comparison of the mean imputation approach with a correctly specified GRE model and a misspecified FM model when estimating rater-specific sensitivity and specificity. Equation (5) was used to impute data. Data are generated under a GRE model where $P_d = 0.2$, $\sigma_0 = \sigma_1 = 1.5$, $I = 1000$, and $J = 4$. Further, the sensitivities for the four raters are 0.80, 0.85, 0.90, and 0.95, and the specificities are 0.95, 0.90, 0.85, and 0.80. The verification is at random where $r_0 = r_5 = 0.05$ and $r_x = 0.40$ for $x = 1, 2, 3,$ and 4. 1000 resampled datasets were obtained and means (SE) across these datasets are presented.

Rater		Truth	MI	SPE	RW	GRE	FM
1	$SENS$	0.80	0.80	0.80	0.79	0.80	0.82
			(0.06)	(0.06)	(0.09)	(0.05)	(0.04)
	$SPEC$	0.95	0.95	0.95	0.95	0.95	0.97
			(0.02)	(0.02)	(0.03)	(0.02)	(0.01)
2	$SENS$	0.85	0.85	0.85	0.84	0.85	0.88
			(0.05)	(0.05)	(0.07)	(0.04)	(0.03)
	$SPEC$	0.90	0.90	0.90	0.90	0.90	0.92
			(0.02)	(0.02)	(0.03)	(0.02)	(0.01)
3	$SENS$	0.90	0.90	0.89	0.90	0.90	0.93
			(0.05)	(0.05)	(0.06)	(0.04)	(0.02)
	$SPEC$	0.85	0.85	0.85	0.85	0.85	0.87
			(0.02)	(0.02)	(0.03)	(0.02)	(0.01)
4	$SENS$	0.95	0.95	0.95	0.95	0.95	0.97
			(0.04)	(0.05)	(0.05)	(0.04)	(0.02)
	$SPEC$	0.80	0.80	0.80	0.80	0.80	0.82
			(0.02)	(0.02)	(0.03)	(0.02)	(0.02)
	P_d	0.20	0.20	0.20	0.20	0.20	0.21
			(0.02)	(0.03)	(0.06)	(0.02)	(0.02)

Asymptotic relative bias of the MI estimator of Prevalence




Asymptotic relative bias under verification biased sampling
Asymptotic relative bias
0.2
0.1
0.0
-0.1
-0.2
2
4
6
8
10
Number of tests J


Asymp. bias under verification biased sampling: weighted imputation
Asymptotic relative bias
0.2
0.1
0.0
-0.1
-0.2
2
4
6
8
10
Number of tests J