5-11-2005

Multiple Imputation for Correcting Verification Bias

Ofer Harel
University of Washington, oharel@stat.uconn.edu

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

1 Introduction

Diagnostic tests seek to distinguish between subjects with a condition to those without. If the true status was known or can be measured by a golden standard procedure, the accuracy of the diagnostic tests could be measured. When considering a binary test (i.e. either positive or negative), one can summarize the results by looking at two measures. One is sensitivity, which is the probability of a positive test given the true status is positive, another one is specificity, which is the probability for a negative test when the true status is negative. When the interest falls on the confidence intervals (CI), finding the CI's for the sensitivity and specificity is equivalent to finding CI's for two Binomial proportions. The commonly used Wald type interval has several limitations due to the nature of the Binomial distribution (Brown et al., 2001). Several alternative intervals have been proposed, and have been proved to be much better than the Wald interval (Brown et al., 2001).

Often, due to various reasons, not all subject's disease statuses can be verified. For example, if true status can be verified only by intrusive operation, the true status for only those who have positive test results would be likely to be verified. In most cases, using only the verified sample, it might cause bias (Zhou et al., 2002). This bias is known as verification bias. Under verification bias we can no longer estimate the sensitivity and specificity using two separated binomial distributions as before. The most widely used correction method for this method was developed by Begg and Greenes (1983) under the ignorable verification bias assumption, that assumes the reason for selecting a sample for verification depends only on observed data. Zhou (1993) extended that method using a maximum likelihood approach. Kosinski and Barnhart (2003) suggested a method for correcting for nonignorable verification bias. Zhou et al. (2002) and Pepe (2003) give a good summary about this subject. The coverage accuracy of the existing methods can be very poor, yielding coverage probabilities much lower than the nominal levels. The reason is that these methods are still using the Wald-type idea for CI. As it is mentioned for the Binomial case, the Wald-type interval does not work.

Estimating the sensitivity, specificity, and their CI's using better alternative methods to Wald-type intervals requires a complete data set (i.e. both test results and true status for all subjects); when there are some subjects without true status, we can not use these better alternative methods directly. By considering verification bias as a missing-data problem, we may use Multiple Imputation (MI) methods for dealing with the missing data problem, which allow us to use better alternative methods for binomial proportions. Multiple imputation (MI) (Rubin, 1987) is a simulated based technique, replacing the missing values with m sets of plausible values, resulting in m sets of "complete" data sets. For each "complete" data set, we compute the sensitivity and specificity estimates and their standard errors and combine them by simple arithmetic rules, giving a valid result taking into consideration the missing values. Using this method allows us not only to use the most common and simplest procedures to estimate the sensitivity, specificity, and their CI's but also gives us a ground to compare

Table 1: Data summary

a. aggregated data

		$T=1$	$T=0$
$V=1$	$D=1$	x^A_{11}	x^A_{10}
	$D=0$	x^A_{01}	x^A_{00}
$V=0$		x^B_{+1}	x^B_{+0}
Total		n_1	n_2

b. complete data

	$T=1$	$T=0$
$D=1$	x_{11}	x_{10}
$D=0$	x_{01}	x_{00}
Total	n_1	n_2

between the different estimation procedures. Our simulation results show that the imputation procedures gives much better results than the currently available procedures. We will be able to pin point one MI method that is the best in coverage accuracy and comparable in interval length and relative bias.

In the remaining parts of this article, we will present the data set up and existing methods in Section 2. In Section 3 we will introduce the use of MI (Rubin, 1987) to address verification bias. We will compare the various techniques using a simulation study in Section 4, give applications of the propose methods to two real data examples in Section 5, and conclude the paper with a discussion in Section 6.

2 Data set up and existing methods

2.1 Data set up

Let T be a binary random variable, indicating whether or not the test was positive ($T=1$) or negative ($T=0$). Since not all subjects' tests are being verified using the golden standard procedure, let V be a random variable indicating whether or not the subject was verified using the golden standard procedure ($V=1$ if verified, $V=0$ if not). Let D be the true status for those who were verified using the golden standard, such that $D=1$ if diseased and $D=0$ if non-diseased (we assume there is no measurement error for the golden standard procedure). Consider Table 1a as a summary of aggregated representation for the data, where the $x's$ are the counts of observations in each status. One can consider $V=0$ to be the indicator for missing data.

We can separate the data into two parts. First, when both T and D are observed ($V=1$), we can call it part A. Second, when T is observed but D is missing ($V=0$), we will refer to this as part B (Table 1a). Each complete data count x_{ij} is a sum of two parts, $x_{ij}=x^A_{ij}+x^B_{ij}$. Although x^A_{ij} is totally observed, x^B_{ij} is not, instead we observe only the marginal total $x^B_{+j}=x^B_{1j}+x^B_{0j}$. The observed data, $Y_{obs}=\{x^A_{ij},x^B_{+j}:i,j=0,1\}$, is represented in Table 1a .

Consider the perfect scenario in which all subjects were verified, and we would have complete data (Table 1b). Even in that case, estimating the specificity and sensitivity might not be a straightforward task. This estimation is equivalent for estimating a proportion from a binomial distribution. Although this problem is considered one of the most basic tasks in elementary statis-

tics, the nature of the binomial distribution makes it less trivial for estimation. Brown et al. (2001) gave a detailed overview of this issue.

2.2 Existing methods

2.2.1 Complete data

When both diagnostic test and true status are available for all subjects, the estimation of the sensitivity and specificity confidence intervals are equivalent to estimating the confidence interval of a binomial proportion.

Estimating a confidence interval for a binomial proportion is a basic issue in statistics. This estimation is not trivial due to the skewed nature of the binomial distribution, especially when the proportion is close to 0 or 1. Consider a random variable $X \sim Bin(n, p)$, the standard interval for p is the Wald interval in which $\hat{p} \pm \kappa\sqrt{n\hat{p}\hat{q}}$, where $\hat{p} = X/n$, $\hat{q} = 1 - \hat{p}$ and κ is the $(1 - \alpha/2)$ percentile of the standard normal distribution. We follow Brown et al. (2001) in the comparison of binomial intervals for complete data. The following methods are known methods, that were developed in order to get around the Binomial estimation problem.

The Agresti-Coull ($A\&C$) interval: Instead of using the standard estimate for the binomial proportion ($p = X/n$), Agresti and Coull (1998) suggested a different estimate. Let $\tilde{X} = X + \kappa^2/2$ and $\tilde{n} = n + \kappa^2$; hence, $\tilde{p} = \tilde{X}/\tilde{n}$ and $\tilde{q} = 1 - \tilde{p}$. The $100(1 - \alpha)\%$ confidence interval for p will be then:

$$\tilde{p} \pm \kappa\sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}} \tag{1}$$

Wilson interval: Consider using the "true" standard error in the confidence interval estimation instead of the estimated one. In that case we use $\sqrt{\frac{pq}{n}}$ instead $\sqrt{\frac{\hat{p}\hat{q}}{n}}$, which lead to the following confidence interval:

$$\frac{X + \kappa^2/2}{n + \kappa^2} \pm \frac{\kappa\sqrt{n}}{n + \kappa^2}\sqrt{\hat{p}\hat{q} + \frac{\kappa^2}{4n}} \tag{2}$$

This interval was introduce by Wilson (1927).

Jeffreys interval: Using a Bayesian approach, it is well known that for a binomial likelihood, one can use a beta conjugate prior. Jeffreys priors are beta priors, and are considered to be noninformative priors (flat priors). Let p have a prior beta distribution $p \sim Beta(1/2, 1/2)$, and let $X \sim Bin(n, p)$. The posterior distribution of p given the data will be $p|X \sim Beta(X + 1/2, n - X + 1/2)$. The (Bayesian) $100(1 - \alpha)\%$ confidence interval is

$$\begin{aligned}(&max(0, Beta(\alpha/2, X + 1/2, n - X + 1/2)),\\ &\quad min(Beta(1 - \alpha/2, X + 1/2, n - X + 1/2)))\end{aligned} \tag{3}$$

where $Beta(\alpha/2, a_1, a_2)$ is the $(1 - \alpha/2)$ quantile of a Beta distribution with parameters a_1 and a_2.

Logit (Rubin) interval: Rubin and Schenker (1987) suggest to use the confidence interval for $\theta = \theta(p) = logit(p) = log(\frac{p}{1-p})$ under a normal approximation. Using a Bayesian argument with the Jeffreys prior distribution, we can show that the distribution of θ is approximately normal. Therefore, if $\hat{\theta}_X$ is the estimate for θ, it follows that $(\theta - \hat{\theta}_X) \sim N(0, V_X)$ where $-V_X^{-1}$ is the second derivative of the log posterior of θ evaluated at $\hat{\theta}_X$. It follows that $\hat{\theta}_X = logit(\tilde{p})$ where $\tilde{p} = \frac{X+1/2}{n+1}$, with $V_X = [(n+1)\tilde{p}(1-\tilde{p})]^{-1}$. Hence, the $100(1-\alpha)\%$ confidence interval is

$$logit^{-1}\{logit(\tilde{p}) \pm \frac{\kappa}{\sqrt{(n+1)\tilde{p}(1-\tilde{p})}}\} \tag{4}$$

Zhou-Li ($Z\&L$) interval: Zhou and Li (2004) proposed a confidence interval based on the Edgeworth expansion of the logit transformation for the proportion in mind. Zhou and Li (2004) consider the pivotal quantity $T = \sqrt{n\hat{p}\hat{q}}(logit(\hat{p}) - logit(p))$; and using the Edgeworth expansion for its distribution they take into account the third and fourth moments while the standard normal approximation takes into account only the first two moments. In order to correct for the skewness term of the expansion they use the function $g(T) = n^{-1/2}b\hat{\gamma} + T + n^{-1/2}a\hat{\gamma}T^2 + n^{-1}(1/3)(a\hat{\gamma})^2T^3$, where $a = -1/6$, $b = 1/6$, and $\hat{\gamma} = \frac{1-2\hat{p}}{\sqrt{\hat{p}\hat{q}}}$. The $100(1-\alpha)\%$ confidence interval is

$$logit^{-1}\left(log\frac{\hat{p}}{\hat{q}} - \frac{g^{-1}(z_{1-\alpha/2})}{\sqrt{n\hat{p}\hat{q}}}, log\frac{\hat{p}}{\hat{q}} - \frac{g^{-1}(z_{\alpha/2})}{\sqrt{n\hat{p}\hat{q}}}\right), \tag{5}$$

where z_α is the α quantile of standard normal, and $g^{-1}(T) = \frac{\sqrt{n}}{a\hat{\gamma}}[(1+3a\hat{\gamma}(\frac{T}{\sqrt{n}} - \frac{b\hat{\gamma}}{n}))^{1/3} - 1]$. In practice, since the probabilities might be 0 or 1, instead of $\hat{p} = X/n$ we used $\breve{p} = \frac{X+1/2}{n+1}$.

2.2.2 Incomplete data methods for verification bias

For incomplete data sets arisen due to verification bias, the estimation of the sensitivity, specificity and their CIs can not follow estimation of binomial proportions anymore. Begg and Greenes (1983) proposed bias correction methods for estimating the sensitivity and specificity. Consider the data given in Table $1a$ with the sample of size n when we know that a sub-sample n_1 has verified disease status. While for the remaining $n_2 = n - n_1$ subjects, we do not know their true disease status. Existing methods for correcting for verification bias are as follows:

Begg-Greenes ($B\&G$) interval: Begg and Greenes (1983) proposed a method to derive sensitivity and specificity under the ignorability assumption for the verification process. If we follow the notation of Table $1a$, it follows that the sensitivity estimate is

$$\hat{\pi}_{1BG} = \frac{(x_{11}^A n_1)/(x_{11}^A + x_{01}^A)}{(x_{11}^A n_1)/(x_{11}^A + x_{01}^A) + (x_{10}^A n_2)/(x_{10}^A + x_{00}^A)},$$

with variance

$$\hat{var}(\hat{\pi}_{1BG}) = (\hat{\pi}_{1BG}(1-\hat{\pi}_{1BG}))^2(\frac{n}{n_1 n_2} + \frac{x^A_{01}}{x^A_{11}(x^A_{11}+x^A_{01})} + \frac{x^A_{00}}{x^A_{10}(x^A_{10}+x^A_{00})}),$$

and the specificity estimate is

$$\hat{\pi}_{2BG} = \frac{(x^A_{00}n_2)/(x^A_{10}+x^A_{00})}{(x^A_{01}n_1)/(x^A_{11}+x^A_{01}) + (x^A_{00}n_2)/(x^A_{10}+x^A_{00})},$$

with variance

$$\hat{var}(\hat{\pi}_{2BG}) = (\hat{\pi}_{2BG}(1-\hat{\pi}_{2BG}))^2(\frac{n}{n_1 n_2} + \frac{x^A_{11}}{x^A_{01}(x^A_{11}+x^A_{01})} + \frac{x^A_{10}}{x^A_{00}(x^A_{10}+x^A_{00})}).$$

Using this information, the $100(1-\alpha)\%$ confidence intervals for sensitivity and specificity will be

$$\hat{\pi}_{1BG} \pm \kappa\sqrt{\hat{var}(\hat{\pi}_{1BG})}, \quad (6)$$

$$\hat{\pi}_{2BG} \pm \kappa\sqrt{\hat{var}(\hat{\pi}_{2BG})}, \quad (7)$$

respectively.

Logit Begg-Greenes interval: Instead of assuming normality for $(\hat{\pi}-\pi)$, one may think that the *logit* transformation of π is closer to a normal approximation, such that $logit(\hat{\pi}) - logit(\pi) \sim N(0, \hat{Var}(logit(\hat{\pi})))$. Using this logit transformation, the $100(1-\alpha)\%$ confidence interval for sensitivity and specificity will be

$$logit^{-1}\left(logit(\hat{\pi}_{1BG}) \pm \kappa\sqrt{\hat{Var}(logit(\hat{\pi}_{1BG}))}\right), \quad (8)$$

$$logit^{-1}\left(logit(\hat{\pi}_{2BG}) \pm \kappa\sqrt{\hat{Var}(logit(\hat{\pi}_{2BG}))}\right), \quad (9)$$

respectively, where $logit(\pi) = log(\frac{\pi}{1-\pi})$,

$$\hat{var}(logit(\hat{\pi}_{1BG})) = (\frac{n}{n_1 n_2} + \frac{x^A_{01}}{x^A_{11}(x^A_{11}+x^A_{01})} + \frac{x^A_{00}}{x^A_{10}(x^A_{10}+x^A_{00})}),$$

and

$$\hat{var}(logit(\hat{\pi}_{2BG})) = (\frac{n}{n_1 n_2} + \frac{x^A_{11}}{x^A_{01}(x^A_{11}+x^A_{01})} + \frac{x^A_{10}}{x^A_{00}(x^A_{10}+x^A_{00})}).$$

Bootstrap method: Instead of using κ, the normal percentile, in the confidence interval, one can use bootstrap-t (Efron and Tibshirani, 1993, ch. 12). Using this method, one is using the bootstrapped values instead of the normal percentile as it was recommended by Pepe (2003).

3 Framework for MI for estimating sensitivity and specificity in the presence of verification bias

When some subjects are not being verified, using multiple imputation will allow us to still use improved complete-data methods for the estimation of the sensitivity, specificity, and their confidence intervals. Multiple imputation (MI) (Rubin (1987); Rubin (1996); Schafer (1997)) is a simulation technique to deal with missing data. We replace each missing value by $m > 1$ plausible values, yielding m complete data sets that differ only in the imputed values. Analyzing each data set by a complete-data method described in Section 2.2.1, will result in m sets of point estimates and standard errors. Combining the results by simple arithmetic rules will provide final estimates and standard errors taking into account the missing data.

In order for the MI to yield valid inference, the simulated values must possess certain properties. MI drawn from a distribution with these qualities was called by Rubin (1987) proper. The full mathematical definition of proper MI is given by Rubin, (1987 pp.118–119). Let Q and U be the population quantity of interest and its variance respectively, and let $\hat{Q}$ be its estimate. We assume that the data can be separated into X, all observed covariates, and $Y = (Y_{obs}, Y_{mis})$, observed and missing values. Since $\hat{Q}$ and U can be created using the imputed Y_{mis} together with the Y_{obs}, one need the estimates from the imputed data sets to be unbiased for Q. For $j = 1, \ldots, m$ imputations, the large–m averages will be $E(\bar{Q}_\infty|X, Y) \doteq \hat{Q}$ and $E(\bar{U}_\infty|X, Y) \doteq U$ as m tends to infinity, while the between imputation variance will be $E(B_\infty|X, Y) \doteq Var(\bar{Q}_\infty|X, Y)$ for large m. Rubin (1987) derives the procedure by Bayesian arguments. However, despite the Bayesian derivation, it has been shown that the method leads to inferences that are well calibrated from frequentists standpoint (Rubin and Schenker, 1986, Schenker and Welsh, 1988, Rubin, 1996, Schafer, 1997).

Schafer (1997) relaxed the *proper* concept to *Bayesianly proper,* where he defined MI to be Bayesianly proper if the imputations are independent realizations of $P(Y_{mis}|Y_{obs})$ when the missingness process can be ignored, or $P(Y_{mis}|Y_{obs}, R)$ when the missingness process can not be ignored. Therefore, Bayesianly proper MI reflects the uncertainty about the missing values (Y_{mis}), given the parameters of the complete-data model.

In addition, Meng (1994) introduced the term congeniality. This term came to relate the Bayesian world and the frequentists world. A model will be called uncongenial if the imputer model and the analysis model differ. More mathematically rigorous definition is in (Meng, 1994).

When the model is congenial and proper, we would get valid inference. If the model is not proper or uncongenial, we will get valid inference only part of the time, depending on the specific scenario. In the next section, we propose a proper MI procedure for correcting for verification bias.

3.1 Imputation stage

Throughout the imputation procedure we use data augmentation (Tanner and Wong, 1987) for imputing the missing values. The main step of MI is to derive the posterior distribution for those with true status, given they were not verified (either positive or negative test). Under ignorability assumption and the structure of the data (Table 1a) one can look at the data as if it came from a multinomial distribution. We can use the multinomial property, in which a conditional multinomial is a multinomial as well (see Appendix 1), to derive the predictive distribution of missing data given the observed data, which is given as follows:

$$(x^B_{1j}, x^B_{0j})|Y_{obs}, \theta \sim M(x^B_{+j}, (\theta_{1j}/\theta_{+j}, \theta_{0j}/\theta_{+j})), \;\; j = 0, 1,$$

where θ_{ij} is the probability that a unit falls into cell (i, j), $\theta_{+j} = \sum_i \theta_{ij}$, and $M(.,.)$ represents a multinomial distribution.

When choosing a Dirichlet prior distribution for multinomial parameters, we obtain the following results which are well known from the conjugate families idea in Bayesian statistics (see Appendix 2).

$$x|\theta \sim M(n, \theta) \tag{10}$$

$$\theta \sim D(\alpha) \tag{11}$$

$$\theta|Y \sim D(\alpha') \tag{12}$$

where $\alpha' = \alpha + x$, and $D(\alpha)$ is a Dirichlet distribution with parameter α.

The data augmentation procedure is drawing iteratively from two distributions. First, one should draw the $x's$ from a multinomial distribution (10), this is done under the assumption that θ is known. Then given those $x's$ values, one should draw values for θ from the (Dirichlet, Beta) posterior distribution (12). The imputation can be carried forward easily using any MI software which allows categorical or loglinear models. For example, SAS (SAS Institute Inc., 1999), and Splus (Schimert et al., 2001). The computational details can be found in Schafer (1997).

The scheme for the imputation stage follows proper imputation draws. Schafer (1997) elaborates on the properties of this model. The use of Jeffreys prior is a common practice in Bayesian analysis when one wants to use a non informative prior (Kass and Wasserman, 1996).

3.2 Analysis stage

After imputing the data, we obtain m sets of complete data sets. Using complete-data methods outlined in Section 2.2.1 we obtain the estimates $(\hat{Q}^1, \hat{Q}^2, \ldots, \hat{Q}^m)$ and associated variances $(U^1, U^2, \ldots, U^m)$ for the sensitivity and specificity. The complete-data methods we are going to use are: Agresti-Coull (A&C) (Agresti and Coull, 1998); Wilson (Wilson, 1927); Jeffrey (Brown et al., 2001); Logit (Rubin) (Rubin and Schenker, 1987); Zhou-Li (Z&L) (Zhou and Li, 2004).

3.3 Combining results

After having m sets of estimates and variances, we use Rubin (1987) combing rules as follows: The overall estimate is $\bar{Q} = \frac{1}{m}\sum Q^i$, $i = 1, \ldots, m$, and its variance is $T = \bar{U} + \frac{1}{m+1}B$, where $\bar{U} = \frac{1}{m}\sum U^i$ is the complete-data variance estimate, and $\frac{1}{m+1}B$ is the variance addition due to the imputations of missing values. The inferences are based on the t-distribution approximation $T^{-1/2}(Q - \hat{Q}) \sim t_\nu$ where the degrees of freedom are $\nu = (m-1)[1+\frac{\bar{U}}{(1+m^{-1})B}]^2$. Therefore, the $100(1-\alpha)\%$ confidence interval for the estimate will be $\bar{Q} \pm t_{\nu,1-\alpha/2}\sqrt{T}$.

4 Simulation study

All competing methods assume large samples. While the $B\&G$ procedures rely on asymptotic results, the MI procedures assume normality which becomes reasonable as the sample size is large. In order to compare the different estimation methods we run two sets of simulation studies. We compare the estimates for the sensitivity and specificity in term of relative bias and the corresponding confidence intervals in terms of interval length and true coverage. The relative bias is calculated as follows: (estimate-true value)/(true value). The first set of simulations are a general scenario while the second set of simulations are based on a real data example (described in the next section). The settings of the first simulation study are as follows: sample sizes of $N = (50, 100, 200)$ represent small, moderate and big sample sizes. The specificity is set up at $Sp = 0.8$, while the sensitivity $Se = (0.9, 0.95)$. The probability of verification given a positive test result is $\lambda_1 = P(V = 1|T = 1) = 0.8$, while the probability of verification given a negative test result is $\lambda_0 = P(V = 1|T = 0) = 0.4$. And let p, the prevalence of the population, be 0.4. We run this simulation 10000 times. For our MI procedures we take $m = 10$, using S-plus 6.2 (Schimert et al., 2001) with a flat (noninformative) prior. The results are summarized in Tables 2–3.

When we tried to use the bootstrap-t method for the $B\&G$ (Begg and Greenes, 1983) and Logit $B\&G$ methods, there was a zero cell in many of the simulated data sets and the estimate of the pivotal quantity was close to zero as well. Hence, the use of this method resulted in very unstable results. We decided not to report them in the comparison tables.

Among the alternative MI techniques compared here, all but the $A\&C$ have relatively similar point estimates. For example, in Table 2 relative biases of the $A\&C$ yield the range from -9% to -5% for the sensitivity and specificity when $N = 50$, and decrease to the range from -3% to -1% for $N = 200$. Relative biases of other MI procedures range from -4% to -2% at $N = 50$, reducing to -1% to -0.5% for $N = 200$. The relative bias of the Rubin (Logit) procedure is ranging from -3% to -0.7% for the sensitivity, and from -1.6% to -0.4% for the specificity. The $B\&G$ and Logit $B\&G$ have the same relative bias which goes from 0.7% to 0.2% for the sensitivity, and 0.3% to -0.1% for the specificity.

Table 2: Simulation results comparing five MI methods and two existing methods for sensitivity (Se) and specificity (Sp) where true values are $Se = 0.9$, $Sp = 0.8$, and 95% coverage. Sample sizes are: a. N=50, b. N=100 c. N=200

a. $N = 50$

	Multiple Imputation						
	A&C	Rubin (Logit)	Wilson	Jeffrey	Z&L	B&G	Logit B&G
Estimated Se	0.817	0.872	0.880	0.880	0.861	0.906	0.906
Estimated Sp	0.761	0.787	0.796	0.796	0.786	0.802	0.802
Se relative Bias	-0.092	-0.031	-0.022	-0.022	-0.043	0.007	0.007
Sp relative Bias	-0.049	-0.016	-0.005	-0.005	-0.018	0.003	0.003
Se coverage	96	94	84	87	78	56	56
Sp coverage	96	96	93	93	97	96	100
Se CI length	0.378	0.463	0.272	0.263	0.292	0.583	0.573
Sp CI length	0.310	0.316	0.277	0.276	0.321	0.621	0.604

b. $N = 100$

	Multiple Imputation						
	A&C	Rubin (Logit)	Wilson	Jeffrey	Z&L	B&G	Logit B&G
Estimated Se	0.854	0.877	0.888	0.888	0.879	0.903	0.903
Estimated Sp	0.781	0.795	0.799	0.799	0.794	0.802	0.802
Se relative Bias	-0.051	-0.026	-0.013	-0.013	-0.023	0.003	0.003
Sp relative Bias	-0.024	-0.006	-0.001	-0.001	-0.008	0.003	0.003
Se coverage	97	95	84	80	83	79	80
Sp coverage	95	96	93	93	96	99	100
Se CI length	0.277	0.330	0.188	0.184	0.196	0.307	0.329
Sp CI length	0.220	0.222	0.199	0.199	0.226	0.448	0.447

c. $N = 200$

	Multiple Imputation						
	A&C	Rubin (Logit)	Wilson	Jeffrey	Z&L	B&G	Logit B&G
Estimated Se	0.876	0.894	0.895	0.895	0.890	0.902	0.902
Estimated Sp	0.789	0.797	0.799	0.799	0.796	0.799	0.799
Se relative Bias	-0.027	-0.007	-0.006	-0.006	-0.011	0.002	0.002
Sp relative Bias	-0.014	-0.004	-0.001	-0.001	-0.005	-0.001	-0.001
Se coverage	93	95	80	79	80	82	86
Sp coverage	95	95	93	93	96	100	100
Se CI length	0.198	0.222	0.131	0.129	0.133	0.185	0.196
Sp CI length	0.156	0.157	0.142	0.142	0.161	0.322	0.323

Table 3: Simulation results comparing five MI methods and two existing methods for sensitivity (Se) and specificity (Sp) where true values are $Se = 0.95$, $Sp = 0.8$, and 95% coverage. Sample sizes are: a. N=50, b. N=100 c. N=200

a. $N = 50$

	Multiple Imputation						
	A&C	Rubin (Logit)	Wilson	Jeffrey	Z&L	B&G	Logit B&G
Estimated Se	0.855	0.915	0.924	0.924	0.903	0.954	0.954
Estimated Sp	0.760	0.786	0.795	0.795	0.785	0.800	0.800
Se relative Bias	-0.1	-0.0367	-0.027	-0.027	-0.050	0.004	0.004
Sp relative Bias	-0.050	-0.018	-0.006	-0.006	-0.019	0	0
Se coverage	97	93	80	85	70	33	32
Sp coverage	95	96	93	92	96	96	99
Se CI length	0.343	0.450	0.240	0.225	0.262	0.482	0.511
Sp CI length	0.308	0.315	0.277	0.277	0.322	0.597	0.583

b. $N = 100$

	Multiple Imputation						
	A&C	Rubin (Logit)	Wilson	Jeffrey	Z&L	B&G	Logit B&G
Estimated Se	0.896	0.932	0.935	0.935	0.924	0.952	0.952
Estimated Sp	0.779	0.793	0.797	0.797	0.792	0.799	0.799
Se relative Bias	-0.059	-0.019	-0.016	-0.016	-0.027	0.002	0.002
Sp relative Bias	-0.026	-0.009	-0.004	-0.004	-0.010	-0.001	-0.001
Se coverage	97	94	82	85	100	55	53
Sp coverage	95	95	93	93	96	99	100
Se CI length	0.235	0.310	0.156	0.147	0.185	0.223	0.253
Sp CI length	0.219	0.222	0.200	0.200	0.227	0.433	0.432

c. $N = 200$

	Multiple Imputation						
	A&C	Rubin (Logit)	Wilson	Jeffrey	Z&L	B&G	Logit B&G
Estimated Se	0.922	0.942	0.943	0.943	0.937	0.951	0.951
Estimated Sp	0.789	0.797	0.799	0.799	0.796	0.800	0.800
Se relative Bias	-0.029	-0.008	-0.007	-0.007	-0.014	0.001	0.001
Sp relative Bias	-0.014	-0.004	-0.001	-0.001	-0.005	0	0
Se coverage	98	94	83	78	81	78	77
Sp coverage	95	95	93	93	96	99	100
Se CI length	0.158	0.195	0.102	0.098	0.106	0.112	0.123
Sp CI length	0.155	0.156	0.142	0.142	0.160	0.309	0.310

Similar results are summarized in Table 3 for the case in which the sensitivity is 0.95.

The MI procedures have advantages over the $B\&G$ methods in both the coverage probability and the interval length. For example, in Table 3 the coverage probabilities of the $B\&G$ methods for sensitivity and specificity are approximately 30% and $96-99\%$ respectively while, the coverage probabilities for the MI Rubin (Logit) procedure for sensitivity and specificity are 93% and 96% respectively, for the small sample size ($N = 50$). For $N = 200$, the coverage probabilities of the $B\&G$ methods for sensitivity and specificity are approximately 78% and 99% respectively, while for the MI Rubin (Logit) procedure the coverage probabilities are 94% and 95% respectively. Another advantage of the MI procedures relative to the existing methods is the CI length. For example, in Table 2a ($N = 50$) the CI length for the specificity of the $B\&G$ methods are 0.621 and 0.604 ($B\&G$ and Logit $B\&G$ respectively), while the length for the MI procedures is $0.276-0.321$ (0.316 for the MI Rubin (Logit) procedure). As the sample size get bigger, the CI get smaller such that for $N = 200$ the CI length for the specificity of the $B\&G$ methods are 0.322 and 0.323 while the length for the MI procedures are $0.142-0.161$ (0.157 for the MI Rubin (Logit) procedure).

The second simulation study mimics the Diaphanography data for breast cancer introduced by Marshall et al. (1981), and used by Greens and Begg (1985) as an illustration of a verification bias problem; see Table 7. Greens and Begg (1985) showed that for an example where (55%) subjects who were tested positive where verified, while only (7%) subjects who were tested negative where verified, there might be a massive bias. In that situation they found that the naive sensitivity estimate of 70% reduced to 28% using their correction. We are going to compare the existing methods to the MI procedures. The simulation settings are as follows: sample sizes of $N = (900, 1500, 3000)$, represent the true example sample size, and two additional big sample sizes. The specificity is set up at $Sp = 0.9$, while the sensitivity $Se = (0.3)$. The probability of verification given a positive test result is $\lambda_1 = P(V = 1|T = 1) = 0.55$, while the probability of verification given a negative test result is $\lambda_0 = P(V = 1|T = 0) = 0.06$. And let p, the prevalence of the population, be 0.03. We run this simulation 10000 times. For our MI procedures we take $m = 10$, using S-plus 6.2 (Schimert et al., 2001) with a flat (noninformative) prior. The results are summarized in Table 4.

Among the alternative MI techniques compared here (Table 4), all have relatively similar point estimates. For example, the relative biases for sensitivity of the MI procedures range from 22% to 23% at $N = 900$, going down to 4% to 6% for $N = 3000$. The Rubin (Logit) procedure relative bias is moving from 23% to 3.6% for the sensitivity, and is -0.1% for the specificity. The $B\&G$ and Logit $B\&G$ have the same relative bias which goes from 64% to 17% for the sensitivity, and 0.1% to 0% for the specificity.

In this simulations study, there is a distinct advantage for the MI procedures over the existing methods with respect to relative bias. In addition, another

Table 4: Simulation results comparing five MI methods and two existing methods for sensitivity (Se) and specificity (Sp) where true values are $Se = 0.3$, $Sp = 0.9$, and 95% coverage. Sample sizes are: a. N=900, b. N=1500 c. N=3000

a. $N = 900$

	Multiple Imputation						
	A&C	Rubin (Logit)	Wilson	Jeffrey	Z&L	B&G	Logit B&G
Estimated Se	0.366	0.369	0.370	0.370	0.367	0.491	0.491
Estimated Sp	0.897	0.899	0.899	0.899	0.899	0.899	0.899
Se relative Bias	0.220	0.230	0.233	0.233	0.223	0.637	0.637
Sp relative Bias	-0.003	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
Se coverage	91	96	59	58	55	54	61
Sp coverage	95	95	63	94	95	100	100
Se CI length	0.766	0.685	0.300	0.298	0.310	0.468	0.447
Sp CI length	0.043	0.043	0.040	0.040	0.0420	0.241	0.270

b. $N = 1500$

	Multiple Imputation						
	A&C	Rubin (Logit)	Wilson	Jeffrey	Z&L	B&G	Logit B&G
Estimated Se	0.353	0.352	0.352	0.352	0.352	0.412	0.412
Estimated Sp	0.898	0.899	0.899	0.899	0.899	0.899	0.899
Se relative Bias	0.177	0.173	0.173	0.173	0.173	0.373	0.373
Sp relative Bias	-0.002	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
Se coverage	88	95	53	51	51	70	75
Sp coverage	95	95	93	93	95	100	100
Se CI length	0.661	0.607	0.237	0.234	0.241	0.404	0.387
Sp CI length	0.033	0.033	0.031	0.031	0.033	0.187	0.201

c. $N = 3000$

	Multiple Imputation						
	A&C	Rubin (Logit)	Wilson	Jeffrey	Z&L	B&G	Logit B&G
Estimated Se	0.319	0.311	0.315	0.315	0.316	0.350	0.350
Estimated Sp	0.899	0.899	0.899	0.899	0.899	0.900	0.900
Se relative Bias	0.063	0.037	0.050	0.050	0.053	0.167	0.167
Sp relative Bias	-0.001	-0.001	-0.001	-0.001	-0.001	0	0
Se coverage	92	96	55	53	55	74	77
Sp coverage	96	96	94	94	96	100	100
Se CI length	0.519	0.486	0.173	0.170	0.174	0.301	0.294
Sp CI length	0.023	0.023	0.022	0.022	0.023	0.132	0.137

Table 5: Hepatic scintigraphy data

		$T=1$	$T=0$
$V=1$	$D=1$	231	27
	$D=0$	32	54
$V=0$		166	140
Total		429	221

major advantage of the MI procedures over the $B\&G$ methods is in the coverage probability. For example, in Table 4 the coverage probabilities of the $B\&G$ methods for sensitivity and specificity are ranging from 54% to 74% and are 100% respectively while, the coverage probabilities for the MI Rubin (Logit) procedure for sensitivity and specificity are ranging from 95% to 96% and from 95% to 96% respectively.

The results in Table 2 and Table 3 lead us to conclude that the MI based on the Rubin (Logit) method performs the best. Even with small sample size the relative bias does not exceed 4%, and as the sample size increase the relative biased decrease for less then 1%. In addition, it yields confidence intervals that are all close to the nominal level for both sensitivity and specificity, and confidence interval lengths which are shorter then the existing methods for small sample sizes and are similar length for moderate and large sample size. In addition, the results in Table 4 reiterate these conclusions. In this set of simulations although the CI length is not the shortest, the relative bias is the smallest and and the coverage probabilities are the best.

5 Real data examples

5.1 Hepatic scintigraphy data

Let us consider the data previously analyzed by Drum and Christacopoulos (1969), Zhou (1993), and Zhou et al. (2002) about hepatic scintigraphy for liver disease. Hepatic scintigraphy is an imaging scan procedure to detect liver cancer. In this study some of the patients were refereed to disease verification process–liver pathology–which was considered as a golden standard. The data are summarized in Table 5.

Following the notation of the previous section, our observed data are as follows:

$$Y_{obs} = \{x^A_{11} = 231, x^A_{10} = 27, x^A_{01} = 32, x^A_{00} = 54, x^B_{+1} = 166, x^B_{+0} = 140\}.$$

In order to proceed with the data augmentation algorithm, let us choose the parameter for the prior Dirichlet distribution to be $\alpha = (1.5, 1.5, 1.5, 1.5)$ which

Table 6: Results comparing five MI methods, the $B\&G$ as existing method and naive method of sensitivity and specificity – Hepatic scintigraphy data

Procedure	Sensitivity			Specificity		
	Est	SE	CI	Est	SE	CI
Naive	0.895	0.019	(0.858,0.932)	0.628	0.052	(0.526,0.730)
B&G	0.836	0.024	(0.788,0.884)	0.738	0.039	(0.662,0.815)
A&C	0.869	0.024	(0.820,0.918)	0.672	0.049	(0.571,0.772)
Rubin	0.872	0.024	(0.817,0.912)	0.675	0.051	(0.567,0.797)
Wilson	0.869	0.016	(0.837,0.901)	0.672	0.031	(0.610,0.733)
Jeffrey	0.872	–	(0.838,0.901)	0.675	–	(0.611,0.734)
Z&L	0.871	–	(0,1)	0.674	–	(0,1)

implies Jeffreys prior. Therefore, our predictive distributions are as follows:

$$\begin{aligned}(x^B_{11}, x^B_{01})^{t+1}|Y_{obs}, \theta^t &\sim M(x^B_{+1}, (\theta_{11}/\theta_{+1}, \theta_{01}/\theta_{+1}))\\ (x^B_{10}, x^B_{00})^{t+1}|Y_{obs}, \theta^t &\sim M(x^B_{+0}, (\theta_{10}/\theta_{+0}, \theta_{00}/\theta_{+0}))\end{aligned}$$

$$\theta \sim D(\alpha) \qquad \text{and} \qquad \theta|Y \sim D(x_{11}+0.5, x_{10}+0.5, x_{01}+0.5, x_{00}+0.5)$$

where $x_{ij} = x^A_{ij} + x^B_{ij}$ $i, j = 0, 1$, and t is the number of iteration. Using S-plus 6.2 (Schimert et al., 2001) we use MI ($m = 10$) to compare the five methods described in section 2.2.1, the existing method described in section 2.2.2 and the naive procedure using only the verified results. Table 6 summarizes the results for the sensitivity, specificity, and their confidence intervals.

Notice that for the sensitivity results, it seems that the Naive estimate is overestimating, and the $B\&G$ estimator is under estimating. All other estimates (methods) are quite close to each other. The agreement is up to the hundredth digit. Since the sample size of this example is quite larger with respect to the simulation ($N = 650$), and the fact that we have proper MI, we can assume that the simulated MI results are more representative of the data. On the other hand, for estimating the specificity, it seems as if the $B\&G$ is overestimating, the Naive procedure under estimating, while the other procedures agrees up to the thousandth digit.

5.2 Diaphanography data for breast cancer

Marshall et al. (1981) introduced the Diaphanography as a test for detecting breast cancer; Greens and Begg (1985) used the data as an illustration of a verification bias problem; see Table 7. We follow Greens and Begg (1985) and compare the seven estimation methods using this data. The observed data in this case can be represented by Y_{obs} such that,

$$Y_{obs} = \{x^A_{11} = 26, x^A_{10} = 7, x^A_{01} = 11, x^A_{00} = 44, x^B_{+1} = 30, x^B_{+0} = 782\}.$$

Table 7: Diaphanography data for breast cancer

		$T=1$	$T=0$
$V=1$	$D=1$	26	7
	$D=0$	11	44
$V=0$		30	782
Total		67	833

Table 8: Results comparing five MI methods, the $B\&G$ as existing method and naive method of sensitivity and specificity – Diaphanography data for breast cancer

Procedure	Sensitivity			Specificity		
	Est	SE	CI	Est	SE	CI
Naive	0.788	0.071	(0.649,0.927)	0.800	0.054	(0.694,0.906)
B&G	0.280	0.073	(0.127,0.434)	0.974	0.007	(0.960,0.989)
A&C	0.706	0.073	(0.560,0.852)	0.861	0.049	(0.753,0.970)
Rubin	0.717	0.075	(0.548,0.841)	0.869	0.057	(0.721,0.944)
Wilson	0.706	0.054	(0.601,0.812)	0.862	0.012	(0.839,0.884)
Jeffrey	0.718	–	(0.603,0.815)	0.863	–	(0.839,0.885)
Z&L	0.715	–	(0,1)	0.863	–	(0,1)

For the data augmentation procedure we choose again the Dirichlet prior distribution parameter to be $\alpha = (1.5, 1.5, 1.5, 1.5)$, implying Jeffrey's prior. The distributions of the data augmentation are as follows:,

$$\begin{aligned}(x^B_{11}, x^B_{01})^{t+1}|Y_{obs}, \theta^t &\sim M(x^B_{+1}, (\theta_{11}/\theta_{+1}, \theta_{01}/\theta_{+1}))\\ (x^B_{10}, x^B_{00})^{t+1}|Y_{obs}, \theta^t &\sim M(x^B_{+0}, (\theta_{10}/\theta_{+0}, \theta_{00}/\theta_{+0}))\end{aligned}$$

$$\theta \sim D(\alpha) \qquad \text{and} \qquad \theta|Y \sim D(x_{11}+0.5, x_{10}+0.5, x_{01}+0.5, x_{00}+0.5)$$

where $x_{ij} = x^A_{ij} + x^B_{ij}$ $i, j = 0, 1$, and t is the number of iteration. Using S-plus 6.2 (Schimert et al., 2001) we use MI ($m = 10$) to compare the five methods described in section 2.2.1, the existing method described in section 2.2.2 and the naive procedure which using only the verified results. Table 8 summarizes the results for the sensitivity, specificity and their confidence intervals.

Once again, we note that for the sensitivity results, it seems that the Naive estimate is overestimating, and the $B\&G$ estimator is under estimating (really bad in this scenario). All other estimates (methods) are quite close to each other. The agreement is up to the hundredth digit. On the other hand, for estimating the specificity, it seems as if the $B\&G$ is overestimating, the Naive procedure under estimating, while the other procedures agrees up to the thousandth digit. Since the sample size of this example is quite larger then the first simulation sample size ($N = 900$), based on the second simulation study, and the fact that we have proper MI, we can assume that the simulated MI results are

more representative of the data. In addition, since we have found during our simulation study that the best method to use (when having an incomplete data set), is Rubin (Logit) method. We would recommend to use this method.

6 Discussion

In this paper we proposed a proper multiple imputation (MI) procedure to correct for verification bias. Verification bias is a common problem in medical research, and there are existing methods to deal with it, but as we have showed, our MI method performs much better than the existing methods. We have shown that in some scenarios the existing method can still give grossly biased results, while the MI procedures would be much closer to the true answer. The reason for this phenomenon is that the common used existing methods were build upon the estimation of binomial proportion using a Wald-type confidence interval, and it has been shown that a Wald-type interval can perform poorly. By using the MI technique, one can use better alternative techniques for dealing with the complications that arise in estimating binomial proportions, and apply it for the verification bias problem. In addition, using MI allows applying several methods in the analysis stage and to use a sensitivity analysis very easily.

Throughout our analysis we assumed ignorable missingness. This assumption allows us not to model the distribution of the missingness indicators. If on the other hand, one believes that this assumption is not reasonable, it is not difficult to alter the procedure to allow nonignorable missingness. By altering the imputation stage of the procedure, adding a model for the missingness, one can impute missing true condition statuses using the nonignorable model, while the analysis stage and combining the result will follow exactly the same steps of the ignorable model. It is rarely possible to test whether the missingness is ignorable or not, hence it requires a medical reasoning for choosing one over the other.

This is first paper that explores the possibility of using the MI technique to correct for verification bias in one sample problem. Our results show that there is a great potential for developing the MI technique for correcting for verification bias in other types of problems.

Appendix 1 – Multinomial properties

Let x be a multinomial random variable with parameter θ. By indexing the cells in the contingence table using only one subscript ($d = 1, ..., D$), it follows that

$$x|\theta \sim M(n, \theta)$$

with $\theta = (\theta_1, \theta_2, ..., \theta_D)$, where the probability distribution of x is

$$P(x|\theta) = \frac{n!}{x_1!x_2!\cdots x_D!}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_D^{x_D}$$

Suppose that we collapse two cells of the contingency table, adding the frequencies together, such that we produce new table $x^* = (z, x_3, ..., x_D)$, where $z = x_1 + x_2$.

Result 1 *The distribution of x^* is multinomial such that*

$$x^*|\theta \sim M(n, \theta^*),$$

where $\theta^* = (\xi, \theta_3, ..., \theta_D)$, and $\xi = \theta_1 + \theta_2$.

Proof 1 Let us sum the multinomial probabilities for all the x-vectors consistent with z, such that

$$\begin{aligned}
P(x^*|\theta) &= \sum_{j=0}^{z} P(x_1 = j, x_2 = z - j, x_3, ..., x_D) \\
&= \sum_{j=0}^{z} \frac{n!}{j!(z-j)!x_3!\cdots x_D!}\theta_1^j\theta_2^{z-j}\theta_3^{x_3}\cdots\theta_D^{x_D} \\
&= \frac{n!}{z!x_3!\cdots x_D!}\theta_3^{x_3}\cdots\theta_D^{x_D}\sum_{j=0}^{z}\frac{z!}{j!(z-j)!}\theta_1^j\theta_2^{z-j} \\
&= \frac{n!}{z!x_3!\cdots x_D!}\theta_3^{x_3}\cdots\theta_D^{x_D}(\theta_1+\theta_2)^z
\end{aligned}$$

since $\sum_{j=0}^{z}\frac{z!}{j!(z-j)!}\theta_1^j\theta_2^{z-j} = (\theta_1 + \theta_2)^z$.

Result 2 *The conditional distribution of (x_1, x_2) given z (the sum) is multinomial such that*

$$(x_1, x_2)|z, \theta \sim M(z, (\theta_1/\xi, \theta_2/\xi)).$$

Proof 2 By using result 1 continuously on variables x_3 to x_D, those cells will collapse to a single cell such that $x_3 + \cdots + x_D = n - z$. Therefore,

$$\begin{aligned}
(x_1, x_2, n-z)|\theta &\sim M(n, (\theta_1, \theta_2, 1-\xi)) \\
(z, n-z)|\theta &\sim M(n, (\xi, 1-\xi)).
\end{aligned}$$

By the definition of conditional probability, it follows that

$$P(x_1, x_2|z, \theta) = \frac{P(x_1, x_2, z|\theta)}{P(z, n-z|\theta)} = \frac{P(x_1, x_2, n-z|\theta)}{P(z, n-z|\theta)},$$

Since both numerator and denominator are multinomial distributions, we can replace the expressions on the right hand side to get

$$\left[\frac{n!}{x_1!x_2!(n-z)!}\theta_1^{x_1}\theta_2^{x_2}(1-\xi)^{n-z}\right]\left[\frac{n!}{z!(n-z)!}\xi^z(1-\xi)^{n-z}\right]^{-1}$$

which can be reduced to

$$P(x_1, x_2|z, \theta) = \frac{z!}{x_1!, x_2!}(\frac{\theta_1}{\xi})^{x_1}(\frac{\theta_2}{\xi})^{x_2},$$

the desired result.

Although the results are stated such that the collapsing is of two cells, the results are true for any arbitrary sets of collapsing.

Appendix 2 – Dirichlet prior

Let $\theta = (\theta_1, \theta_2, ..., \theta_D)$ be a set of random variables such that $\theta_d \geq 0$ for $d = 1, 2, ..., D$ and $\sum_{d=1}^{D} \theta_d = 1$. The density function of θ given the parameter $\alpha = (\alpha_1, \alpha_2, ...\alpha_D)$, is

$$P(\theta|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_D)}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\cdots\theta_D^{\alpha_D-1}$$

where $\alpha_0 = \sum_{d=1}^{D}\alpha_d$ and $\Gamma(.)$ denotes the gamma function. This Dirichlet distribution is often written as $\theta|\alpha \sim D(\alpha)$. When used as a prior for a multinomial distribution, it is typical to omit the normalizing constant such that,

$$\pi(\theta) \propto \theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\cdots\theta_D^{\alpha_D-1}$$

where $(\alpha_1, ..., \alpha_D)$ are user specific hyperparameters. Since the likelihood function of a multinomial distribution is

$$L_{x|\theta} = \frac{n!}{x_1x_2!\cdots x_D!}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_D^{x_D},$$

the posterior distribution is the product of the prior function (information) and the likelihood function, leading us to

$$\begin{aligned} L_{\theta|x} = \pi(\theta) \times L_{x|\theta} &\propto K \times (\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\cdots\theta_D^{\alpha_D-1})(\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_D^{x_D}) \\ &= K \times \theta_1^{x_1+\alpha_1-1}\theta_2^{x_2+\alpha_2-1}\cdots\theta_D^{x_D+\alpha_D-1} \\ &\sim D(x+\alpha), \end{aligned}$$

a Dirichlet posterior distribution with parameter $(x + \alpha) = (x_1 + \alpha_1, x_2 + \alpha_2, ..., x_D + \alpha_D)$.

References

Agresti, A. and B. A. Coull (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician 52*, 119–126.

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

Brown, L. D., T. T. Cai, and A. Dasgupta (2001). Interval estimation for a binomial proportion (Pkg: p101-133). *Statistical Science 16*(2), 101–133.

Drum, D. and J. Christacopoulos (1969). Hepatic scintigraphy in clinical decision making. *Journal of Nuclear Medicine 13*, 908–915.

Efron, B. and R. Tibshirani (1993). *An introduction to the bootstrap*. Chapman & Hall Ltd.

Greens, R. and C. Begg (1985). Assessment of diagnostic technologies: Methodology for unbiased estimation from samples of selective verified patients. *Investigative Radiology 20*, 751–756.

Kass, R. E. and L. Wasserman (1996). The selection of prior distributions by formal rules (Corr: 1998V93 p412). *Journal of the American Statistical Association 91*, 1343–1370.

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

Marshall, V., W. D. C, and S. K. D (1981). Diaphanography as a means of detecting breast cancer. *Radioloy 150*, 339–343.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (Disc: p558-573). *Statistical Science 9*, 538–558.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association 91*, 473–489.

Rubin, D. B. and N. Schenker (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association 81*, 366–374.

Rubin, D. B. and N. Schenker (1987). Logit-based interval estimation for binomial data using the Jeffreys prior. *Sociological Methodology 0*, 131–144.

SAS Institute Inc. (1999). *SAS Procedure Guide* (version 8 ed.). Cary, NC: SAS Institute Inc.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schenker, N. and A. H. Welsh (1988). Asymptotic results for multiple imputation. *The Annals of Statistics 16*, 1550–1566.

Schimert, J., J. L. Schafer, T. Hesterberg, C. Fraley, and D. Clarkson (2001). *Analyzing Missing Values in S-PLUS*. Seattle, WA: Insightful Corp.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation (C/R: p541-550). *Journal of the American Statistical Association 82*, 528–540.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association 22*, 209–212.

Zhou, X. H. (1993). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communications in Statistics, Part A – Theory and Methods 22*, 3177–3198.

Zhou, X. H. and C. Li (2004). Improving interval estimation of binomial proportions. submitted.

Zhou, X. H., N. A. Obuchowski, and D. M. Obuchowski (2002). *Statistical Methods is Diagnostic Medicine*. New York, USA: Wiley & Sons.