

Correction of Verification Bias using Log-linear Models for a Single Binary-scale Diagnostic Tests

Haresh Rochani, Hani Samawi, Robert Vogel and Jingjing Yin

Department of Biostatistics, Jiann-Ping Hsu College of Public Health, USA

Abstract

In diagnostic medicine, the test that determines the true disease status without an error is referred to as the gold standard. Even when a gold standard exists, it is extremely difficult to verify each patient due to the issues of cost-effectiveness and invasive nature of the procedures. In practice some of the patients with test results are not selected for verification of the disease status which results in verification bias for diagnostic tests. The ability of the diagnostic test to correctly identify the patients with and without the disease can be evaluated by measures such as sensitivity, specificity and predictive values. However, these measures can give biased estimates if we only consider the patients with test results who also underwent the gold standard procedure. The emphasis of this paper is to apply the log-linear model approach to compute the maximum likelihood estimates for sensitivity, specificity and predictive values. We also compare the estimates with Zhou's results and apply this approach to analyze Hepatic Scintigraph data under the assumption of ignorable as well as non-ignorable missing data mechanisms. We demonstrated the efficiency of the estimators by using simulation studies.

Keywords: Verification bias; Diagnostic tests; Log-linear models; Missing data

Introduction

Diagnostic testing in medicine is the process of identifying the patients with and without a particular disease. Accuracy of the diagnostic test is the ability of the test to correctly identify the true disease status of the patient [1]. The test or procedure that determines the true disease status without an error is called the gold standard test. However, even with the existence of a gold standard test, verification for the disease status of each patient may not be obtained due to various reasons such as the invasive nature or too costly gold standard test. For example, the Prostate Specific Antigen (PSA) blood test is used as a screening test for diagnosis of prostate Cancer with a ranging cost from \$60 to \$80. However, the true diagnosis is generally confirmed by invasive procedures such as prostate biopsy with a ranging cost from \$1600 to \$1800. Thus patients with a high risk or a prostate screening test positive are more likely to be offered the gold standard test than those with low risk or a negative screening test. Furthermore, the inference about the measures of diagnostic accuracy such as sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) may be biased because individuals who are selected for the gold standard test based on the diagnostic test results are not a random sample [2]. Many authors have referred to this bias as, work-up bias [3] or verification bias [4]. Therefore, in the presence of verification bias, the efficiency of a diagnostic test heavily depends only on those patients whose disease status has been verified. In addition if there is a positive association between patient selection for verification and the test results will produce more bias results [5].

When the true disease status is missing among patients who were not selected for verification, the framework of the missing data mechanism proposed by Rubin can be applied for verification bias correction. If the probability of selecting patients for verification is independent of both observed and unobserved data then missingness in the disease status is considered missing completely at random (MCAR). Missingness in the disease status is considered missing at random (MAR) when the probability of selecting patients only depends on the observed data and it is considered missing not at random (MNAR) when the probability depends on unobserved data [6]. This is most likely to occur when

there is long time lag between the initial test and verification, when there are multiple investigators at various institutions, when the patient population is very heterogeneous or when the disease process is not well understood.

Begg and Greenes [4] proposed a method of verification bias correction for binary diagnostic tests by using Bayes' theorem under the MAR (i.e. conditional assumption). Zhou [7] derived the maximum likelihood estimators for sensitivity and specificity and their variances under both MAR and MNAR mechanisms. However, under the MNAR assumption, Zhou [7] assumed that two known ratios need to be quantified. Zhou [8] showed that the estimators of PPV and NPV are unbiased and consistent if the probability of selecting patients for disease verification does not depend on the true disease status of the patient (i.e. MAR assumption). Some authors have tried to use multiple imputation methods to correct for verification bias under ignorable missingness [9]. However, the validity of that approach was debated by De Groot et al. [10]. In dealing with multiple diagnostic tests and the presence of covariates, Baker [11] Kosinski and Barnhart [12] suggested regression approaches to deal with the MNAR missing mechanism. However, their models require the use of multiple tests or covariates in order to achieve identifiability. Martinez et al. [13] have tried to address the issue of verification bias in the MNAR setting by using a Bayesian approach with a beta prior distribution. Some authors have tried using Gibbs sampling techniques as a Bayesian approach to correct the verification bias in MNAR estimates [14]. All these available methods require iterative methods to compute the estimates of measures for

***Corresponding author:** Haresh Rochani, Assistant Professor, Department of Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, 30460, USA, Tel: 912.478.1101; E-mail: hrochani@georgiasouthern.edu

Received October 30, 2015; **Accepted** December 11, 2015; **Published** December 18, 2015

Citation: Rochani H, Samawi H, Vogel R, Yin J (2015) Correction of Verification Bias using Log-linear Models for a Single Binary-scale Diagnostic Tests. J Biom Biostat 6: 266. doi:10.4172/2155-6180.1000266

Copyright: © 2015 Rochani H, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

diagnostic accuracy especially under MNAR mechanism. However, by using the log-linear model approach, the explicit solutions for the measures of diagnostic accuracy estimates can be computed by the use of a simple algebraic formula (computationally convenient) under all missing data mechanisms.

In this paper, we provide the log-linear model approach to correct for verification bias in order to compute the estimates of diagnostic measures (i.e. Sensitivity, Specificity, PPV and NPV) under various assumptions of missing data mechanisms. We also confirm by simulations that the estimators are asymptotically unbiased and more efficient than complete cases estimates. In next section, we will show the impact of verification bias by using hypothetical example. The following sections will focus on the method of log-linear models to correct the verification bias for diagnostic measures and comparison with Zhou [7] results. Furthermore, we illustrate the log-linear model approach by application to the Hepatic Scintigraph data.

Impact of Verification Bias

In diagnostic medicine, it is a common practice to refer only patients who are test positives for a gold standard test for disease verification. To illustrate the impact of bias, we considered a hypothetical study of 1000 patients to measure the sensitivity, specificity, PPV and NPV (Tables 1 and 2).

The hypothetical data are summarized in Table 1. If every patient is being referred for verification for the gold standard then the true sensitivity from Table 1 is 0.71. Furthermore, the true specificity, PPV and NPV are 0.87, 0.37 and 0.96 respectively. However, if we consider that only 10% of those who have a test negative result receive the gold standard for verification, then the data are summarized in Table 2. From Table 2, the observed sensitivity, specificity, PPV and NPV are 0.96, 0.4, 0.37 and 0.96 respectively. Overall if positive test results are more likely to receive disease verification, (i.e. MAR) then by using only verified cases, the estimates for sensitivity and specificity are biased, in fact, sensitivity is being overestimated and specificity is being underestimated. However, under the MAR assumption, the estimates of PPV = $P(D+|T+)$ and NPV = $P(D-|T-)$ from complete cases are unbiased [8,15,16].

Methodology

Log-linear parameterization

Baker et al.[17] used the homogeneous log-linear model approach to adjust the expected cell counts for a two way table with missing data. We can use the similar approach (BRD model approach) to compute the estimates for measures of diagnostic accuracy correcting for

	Verified		2*Unverified	2*Total
-3	D +	D -		
T +	71	120	0	191
T -	29	780	0	809
				1000

Table 1: Hypothetical data.

	Verified		2*Unverified	2*Total
-3	D +	D -		
T +	71	120	0	191
T -	3	78	728	809
				1000

Table 2: Hypothetical data with unverified cases.

		Verified		Unverified
		R _{J=1}		R _{J=2}
		J=1	J=2	
		(D +)	(D -)	
R _{I=2}	(T +) I=1	n ₁₁₁₁	n ₁₂₁₁	n ₁₊₁₂
	(T -) I=2	n ₂₁₁₁	n ₂₂₁₁	n ₂₊₁₂
				n ₊₊₁₊

Table 3: Two way table with one supplementary margin.

verification bias. To demonstrate this idea, define two random variables I and J, each with two levels i = 1,2 and j = 1,2. Let R_J be the missing data indicator for variable J, such that R_J = 1 if J is being observed and R_J = 2, if J is missing. Similarly, let R_I = 1 if we observe in variable I. Consider I × J cross-classified Table 3 with total sample size of n₊₊₁₊, each with two levels with one supplemental margin (in the terminology of Dempster and Baker 1992) with cell count n_{ijkl}. The cell counts with R_I = R_J = 1 is denoted as n_{ij11}. The supplemental margin for variable J when R_J = 2 is denoted by n_{i+12}. In these cases, the symbol '+' denotes the summation over the corresponding index. Furthermore, denote the expected cell counts by λ_{ijkl}, cell probabilities for I, J, R_I and R_J by π_{ijkl} and the marginal probabilities by π_{ij.} for Table 3.

Log-linear models are widely used to analyze contingency tables. Missing data indicators for contingency tables with supplemental margins can be incorporated in log-linear models in the same way as they incorporate the other variables. In terms of λ, the homogenous log-linear model (presence of only two-way interactions) for partially observed one variable is $\log(\mu_{ijkl}) = \lambda + \lambda_i(i) + \lambda_j(j) + \lambda_{R_i}(1) + \lambda_{R_j}(k) + \lambda_{ij}(ij) + \lambda_{R_i}(i1) + \lambda_{R_j}(ik)$

$$+ \lambda_{R_i}(j1) + \lambda_{R_j}(jk) + \lambda_{R_i R_j}(1k) \quad (1)$$

Because this is an over parameterized model, we add the side conditions that the sum of each λ term over each of the indicated subscripts is zero. For computational purposes, it is convenient to use an alternative parameterization of the model. Let m_{ij} ≥ 0, b_{ij} ≥ 0 and $\sum_i \sum_j m_{ij}(1+b_{ij}) = n_{++1+}$ such that μ_{ij11} = m_{ij} and μ_{ij12} = m_{ij}b_{ij}. In terms of log-linear models $m_{ij} = \exp\{\lambda + \lambda_i(i) + \lambda_j(j) + \lambda_{R_i}(1) + \lambda_{R_j}(1) + \lambda_{ij}(ij) + \lambda_{R_i}(i1) + \lambda_{R_j}(i1) + \lambda_{R_i}(j1) + \lambda_{R_j}(j1) + \lambda_{R_i R_j}(11)\}$

Taking advantage of the side conditions, we can derive the following expression for b_{ij}:

$$b_{ij} = -2 \exp\{\lambda_{R_i}(1) + \lambda_{R_j}(i1) + \lambda_{R_j}(j1) + \lambda_{R_i R_j}(11)\}$$

The detailed derivation can be found in the Appendix.

Denote,

$$p_1 = Pr(R_I = 1, R_J = 1 | I = i, J = j) = \frac{\mu_{ij11}}{n_{++1+}}$$

$$p_2 = Pr(R_I = 1, R_J = 2 | I = i, J = j) = \frac{\mu_{ij12}}{n_{++1+}}$$

In probabilistic terms, m_{ij} and b_{ij} can be expressed as follows,

$$m_{ij} = n_{++1+} p_1$$

$$b_{ij} = \frac{\mu_{ij12}}{\mu_{ij11}} = \frac{\mu_{ij12} n_{++1+}}{\mu_{ij11} n_{++1+}} = \frac{p_2}{p_1}$$

For Table 3, the total expected count is

$$\mu_{++1+} = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 \mu_{ijkl} = \sum_{i=1}^2 \sum_{j=1}^2 \mu_{ij11} + \sum_{i=1}^2 \sum_{j=1}^2 \mu_{ij12} = \sum_{i=1}^2 \sum_{j=1}^2 m_{ij} (1 + b_{ij})$$

and in general notation the marginal and conditional probabilities can be expressed as :

$$\pi_{ij..} = Pr(I = i, J = j) = \frac{m_{ij}(1 + b_{ij})}{n_{++}}$$

$$\pi_{i+..} = Pr(I = i) = \sum_{j=1}^2 \frac{m_{ij}(1 + b_{ij})}{n_{++}}$$

$$\pi_{+j..} = Pr(J = j) = \sum_{i=1}^2 \frac{m_{ij}(1 + b_{ij})}{n_{++}}$$

$$\pi_{ijj..} = Pr(I = i | J = j) = \frac{\pi_{ij..}}{\pi_{+j..}} = \frac{m_{ij}(1 + b_{ij})}{\sum_{i=1}^2 m_{ij}(1 + b_{ij})}$$

$$\pi_{jii..} = Pr(J = j | I = i) = \frac{\pi_{ij..}}{\pi_{i+..}} = \frac{m_{ij}(1 + b_{ij})}{\sum_{j=1}^2 m_{ij}(1 + b_{ij})}$$

The likelihood ratio statistic with respect to a model fitting the data perfectly is,

$$G^2 = 2 \left\{ \sum_{i,j} n_{ij11} \log \left(\frac{n_{ij11}}{\hat{m}_{ij}} \right) + \sum_i n_{i+12} \log \left(\frac{n_{i+12}}{\sum_j \hat{m}_{ij} b_{ij}} \right) + n_{++12} \log \left(\frac{n_{++12}}{\sum_{i,j} \hat{m}_{ij} b_{ij}} \right) \right\}$$

Models parameter estimates

Baker et al. [17] identified nine different models for the two-way table with three supplementary margins based on the dependence of missingness on one or the other or its own realization. For Table 3, when b_{ij} equals to β_j , β_i or β_{ij} , there are three different identifiable models. First and second subscripts to the parameter β correspond to variable I and J respectively. The subscript "." indicates that parameter is constant over corresponding index. In addition, each model has a unique interpretation. For example, Model (β_j) can be interpreted as the missingness in J is constant (MCAR). Similarly, Model (β_i) can be interpreted as the missingness in J depends on the realization of I (MAR) and when the missingness in J depends on its own realization, then we consider Model β_{ij} to be MNAR.

Assuming Poisson distribution for cell counts, the maximum likelihood estimates for m_{ij} and β can be derived from log-likelihood functions for each model. Detailed derivation of the estimates are provided in Appendix. The likelihood ratio statistics for each model can also be found in the Appendix.

MCAR estimates

$$\hat{\beta}_{..} = \frac{n_{++12}}{n_{++1+}} \quad \hat{m}_{ij} = \frac{n_{ij11} \cdot n_{i+1+} \cdot n_{++11}}{n_{++1+} \cdot n_{i+11}}$$

$$\hat{\pi}_{ijj..} = \frac{\hat{m}_{ij}(1 + \hat{\beta}_{..})}{\sum_{i=1}^2 \hat{m}_{ij}(1 + \hat{\beta}_{..})} = \frac{\hat{m}_{ij}}{\sum_{i=1}^2 \hat{m}_{ij}} \quad (2)$$

$$\hat{\pi}_{jii..} = \frac{\hat{m}_{ij}(1 + \hat{\beta}_{..})}{\sum_{j=1}^2 \hat{m}_{ij}(1 + \hat{\beta}_{..})} = \frac{\hat{m}_{ij}}{\sum_{j=1}^2 \hat{m}_{ij}} \quad (3)$$

From equation 2, the estimates for Sensitivity and Specificity are

$$Sensitivity(MCAR) = \hat{\pi}_{i=1|j=1..} = \frac{\hat{m}_{11}}{\hat{m}_{11} + \hat{m}_{21}}$$

$$Specificity(MCAR) = \hat{\pi}_{i=2|j=2..} = \frac{\hat{m}_{22}}{\hat{m}_{12} + \hat{m}_{22}}$$

Similarly, the estimates for PPV ($\hat{\pi}_{j=1|i=1..}$) and NPV ($\hat{\pi}_{j=2|i=2..}$) can be computed from equation 3.

MAR estimates

$$\hat{\beta}_i = \frac{n_{i+12}}{\hat{m}_{i+}} \quad \hat{m}_{ij} = n_{ij11}$$

$$\hat{\pi}_{ijj..} = \frac{\hat{m}_{ij}(1 + \hat{\beta}_i)}{\sum_{i=1}^2 \hat{m}_{ij}(1 + \hat{\beta}_i)} \quad (4)$$

$$\hat{\pi}_{jii..} = \frac{\hat{m}_{ij}(1 + \hat{\beta}_i)}{\sum_{j=1}^2 \hat{m}_{ij}(1 + \hat{\beta}_i)} \quad (5)$$

From equation 4 the estimates of the Sensitivity and Specificity can be obtained as follows

$$Sensitivity(MAR) = \hat{\pi}_{i=1|j=1..} = \frac{\hat{m}_{11}(1 + \hat{\beta}_1)}{\hat{m}_{11}(1 + \hat{\beta}_1) + \hat{m}_{21}(1 + \hat{\beta}_2)} \quad (6)$$

$$Specificity(MAR) = \hat{\pi}_{i=2|j=2..} = \frac{\hat{m}_{22}(1 + \hat{\beta}_2)}{\hat{m}_{12}(1 + \hat{\beta}_1) + \hat{m}_{22}(1 + \hat{\beta}_2)} \quad (7)$$

can be obtained. Similarly, from equation 5 the estimates of PPV ($\hat{\pi}_{j=1|i=1..}$) and NPV ($\hat{\pi}_{j=2|i=2..}$) can be computed.

MNAR estimates

$$\sum_{j=1}^2 \hat{m}_{ij} \hat{\beta}_j = n_{i+12} \quad \hat{m}_{ij} = n_{ij11}$$

$$\hat{\pi}_{ijj..} = \frac{\hat{m}_{ij}(1 + \hat{\beta}_j)}{\sum_{i=1}^2 \hat{m}_{ij}(1 + \hat{\beta}_j)} \quad (8)$$

$$\hat{\pi}_{jii..} = \frac{\hat{m}_{ij}(1 + \hat{\beta}_j)}{\sum_{j=1}^2 \hat{m}_{ij}(1 + \hat{\beta}_j)} \quad (9)$$

For Model (β_j), equation 8 provides the estimates of Sensitivity and Specificity as follows

$$Sensitivity(MNAR) = \hat{\pi}_{i=1|j=1..} = \frac{\hat{m}_{11}(1 + \hat{\beta}_1)}{\hat{m}_{11}(1 + \hat{\beta}_1) + \hat{m}_{21}(1 + \hat{\beta}_1)}$$

$$Specificity(MNAR) = \hat{\pi}_{i=2|j=2..} = \frac{\hat{m}_{22}(1 + \hat{\beta}_2)}{\hat{m}_{12}(1 + \hat{\beta}_2) + \hat{m}_{22}(1 + \hat{\beta}_2)}$$

while equation 9 provides the estimates of PPV ($\hat{\pi}_{j=1|i=1..}$) and NPV ($\hat{\pi}_{j=2|i=2..}$).

It is possible to obtain negative maximum likelihood estimates for $\hat{\beta}_j$ in MNAR model. If any solution is negative ($\hat{\beta}_1$ or $\hat{\beta}_2 < 0$), the estimates still can be computed by maximizing the likelihood function by using the limited memory algorithm for constrained optimization Byrd [18]. Optim function can be utilized in statistical softwar R 2.15.1 or higher versions.

Inference about diagnostic measures

In this section, the multivariate delta method is used to draw the

inferences about the diagnostic measures for Model (β_i) (MAR). A similar method can be used for Model (β) (MCAR) and Model (β_j) (MNAR). To derive the asymptotic variance of the estimates, we have

$$\frac{\hat{\pi}_{ij.} - \pi_{ij.}}{\sqrt{\widehat{Var}(\hat{\pi}_{ij.})}} \sim \mathcal{N}(0,1).$$

However, to ensure better normal approximation, we derived the asymptotic distribution of

$\text{Logit}(\hat{\pi}_{ij.}) = \log\left[\frac{\hat{\pi}_{ij.}}{1-\hat{\pi}_{ij.}}\right]$. By using multivariate delta method for Model(β_i), we have

$$\text{Logit}(\hat{\pi}_{ij.}) \sim \mathcal{N}(\text{logit}\pi_{ij.}, \mathbf{C}^T \mathbf{I}_2^{-1} \mathbf{C}) \quad (10)$$

where $\mathbf{C}^T = \left(\frac{\partial \text{logit}\pi_{ij.}}{\partial m_{11}}, \frac{\partial \text{logit}\pi_{ij.}}{\partial m_{12}}, \frac{\partial \text{logit}\pi_{ij.}}{\partial m_{21}}, \frac{\partial \text{logit}\pi_{ij.}}{\partial m_{22}}, \frac{\partial \text{logit}\pi_{ij.}}{\partial \beta_1}, \frac{\partial \text{logit}\pi_{ij.}}{\partial \beta_2} \right)$ and

\mathbf{I}_2 is the Fisher information. The detailed derivation of \mathbf{I}_2 is provided in the Appendix. Therefore, the $100(1-\delta)\%$ confidence interval for $\pi_{ij.}$ is

$$\left(\frac{\exp\left(\log\frac{\hat{\pi}_{ij.}}{1-\hat{\pi}_{ij.}} - z_{1-\delta/2} \sqrt{\widehat{Var}\left(\log\frac{\hat{\pi}_{ij.}}{1-\hat{\pi}_{ij.}}\right)}\right)}{1 + \exp\left(\log\frac{\hat{\pi}_{ij.}}{1-\hat{\pi}_{ij.}} - z_{1-\delta/2} \sqrt{\widehat{Var}\left(\log\frac{\hat{\pi}_{ij.}}{1-\hat{\pi}_{ij.}}\right)}\right)}, \frac{\exp\left(\log\frac{\hat{\pi}_{ij.}}{1-\hat{\pi}_{ij.}} + z_{1-\delta/2} \sqrt{\widehat{Var}\left(\log\frac{\hat{\pi}_{ij.}}{1-\hat{\pi}_{ij.}}\right)}\right)}{1 + \exp\left(\log\frac{\hat{\pi}_{ij.}}{1-\hat{\pi}_{ij.}} + z_{1-\delta/2} \sqrt{\widehat{Var}\left(\log\frac{\hat{\pi}_{ij.}}{1-\hat{\pi}_{ij.}}\right)}\right)} \right)$$

In case of negative estimates for MNAR model, the variances for the measures of the diagnostic accuracy can be estimated by bootstrap method.

Comparison with Zhou's result

Zhou [7] derived the estimates for Sensitivity and Specificity under MAR and MNAR assumptions. The estimates obtained in previous section for Sensitivity and Specificity under MAR assumption are equivalent to Zhou's estimates (Appendix). However, under MNAR assumption Zhou [7] assumes that the two ratios

$$e_0 = \frac{P(R_j = 1 | I = 2, J = 1)}{P(R_j = 1 | I = 2, J = 2)} \quad e_1 = \frac{P(R_j = 1 | I = 1, J = 1)}{P(R_j = 1 | I = 1, J = 2)}$$

are known to estimate the Sensitivity and Specificity. However, the log-linear model approach does not require such assumptions to estimate MNAR parameters.

Simulation Study

A simulation study was conducted to illustrate the reduction in bias and root mean squared error (RMSE) of the estimators for sensitivity and specificity for MAR model compared to complete case analysis (Table 4).

Two variables "Test Results" and "Disease Status" with different sample sizes (N) were generated from the multinomial distribution with marginal probabilities of $\hat{\pi}_{+1} = 0.5$ and $\hat{\pi}_{+2} = 0.5$ under different sensitivity ($= \frac{\pi_{11}}{\pi_{+1}}$) and specificity ($= \frac{\pi_{22}}{\pi_{+2}}$) values. We cross-classified these variables in a 2×2 contingency table. This process was repeated 2000 times. In each complete contingency tables, we generated missing values in "Disease Status" in such a way that missingness in "Disease Status" depends on the realization of the "Test Results" to ensure the MAR missing mechanism. Table 4 represents the results from our simulation for MAR model when 50% of negative tests and 10% of the positive tests were not verified for the disease. Table 4 shows that correcting for the verification bias using the log-linear model approach will substantially reduce the bias as well as improve the performance

N	(Sen %, Sp%)	Complete Cases		Model Based	
		Bias (RMSE)		Bias (RMSE)	
		Sen	Sp	Sen	Sp
200	(60,60)	0.1464	0.1269	0.0623	0.0375
		(0.1438)	(0.1279)	(0.0005)	(0.0021)
	(95,95)	0.0504	0.0229	0.0245	0.0195
		(0.0375)	(0.0214)	(0.0004)	(0.0004)
	(60,95)	0.1470	0.0230	0.0586	0.0200
		(0.1424)	(0.0213)	(0.0014)	(0.0004)
500	(95,60)	0.1464	0.1279	0.0623	0.0375
		(0.2235)	(0.3587)	(0.1585)	(0.1921)
	(60,60)	0.1461	0.1310	0.0389	0.0228
		(0.1461)	(0.1310)	(0.0017)	(0.0017)
	(95,95)	0.0396	0.0212	0.0160	0.0122
		(0.0357)	(0.0210)	(0.0002)	(0.0008)
	(60,95)	0.1450	0.0217	0.0354	0.0121
		(0.1450)	(0.0215)	(0.0001)	(0.0001)
	(95,60)	0.0412	0.1301	0.0159	0.0234
		(0.0373)	(0.1301)	(0.0006)	(0.0006)

Table 4: Bias and MSE comparison for 50% missing in T- and 10% missing in T+.

of the estimators compared to complete case analysis. We obtained similar results with 10% missingness in T+ with different amounts of missingness in T- such as 70%, 60%, 40%, 30% and 20% and/or various marginal probabilities. In addition to this, we carried out an extensive simulation for Model (β_j) by simulating the MNAR missing mechanism to estimate PPV and NPV. We observed the reduction in bias and RMSE compared to complete case analysis under different amounts of non-ignorable missingness on the disease status.

Application

In this section, we will demonstrate the log-linear modeling approach to compute the measures of diagnostic accuracy under various missing mechanisms by using the Hepatic Scintigraph data published by [19]. In the Hepatic Scintigraph data, 650 patients underwent hepatic scintigraphy of which 429 patients had positive scans while 221 patients had negative hepatic scans. Only 263 patients out of 429 and 81 patients out of 221 were referred for status verification using procedures such as liver biopsy, exploratory laparotomy or autopsy within 6 weeks of their scans. Table 5 represents the verified and unverified cases by hepatic scan results. Table 6 shows the Maximum likelihood estimates for model parameters (Tables 5-7).

Table 7 represents the estimates of sensitivity, specificity, PPV and NPV and 95% confidence intervals under three different models by using the log-linear model approach to Table 5. By using Zhou [7] method the estimates of Sensitivity and Specificity can range from (0.68,0.95) and (0.37,0.86) respectively depending upon the values of e_0 and e_1 . From Table 7, it can be confirmed that under MAR and MCAR assumptions the estimates of PPV and NPV and under MNAR assumption the estimates of sensitivity and specificity can be obtained from complete cases [8].

Final Remarks

Verification bias is an extremely common problem in diagnostic medicine. This paper shows that how the log-linear model approach in single binary-scale diagnostic tests correct the verification bias for estimating the diagnostic measures. Log-linear models also reduces

	Verified		Unverified	Total
	D+	D-		
T	231	32	166	429
T-	27	54	140	221
				650

Table 5: Hepatic scintigraph data.

Model	$\hat{\beta}$		G ²	P-Value
	$\hat{\beta}_1$	$\hat{\beta}_2$		
Model (β_j)	0.89		35.84	< 0.001
Model (β_i)	0.63	1.73	-	-
Model (β_j)	0.39	2.4	-	-

Table 6: ML estimates for model parameters.

Measure	Model (β_j)	Model (β_i)	Model (β_j)
Sensitivity	0.9	0.84	0.9
	(0.86,0.93)	(0.77,0.88)	(0.85,0.93)
Specificity	0.63	0.74	0.63
	(0.52,0.73)	(0.66,0.81)	(0.52,0.72)
PPV	0.88	0.88	0.75
	(0.83,0.91)	(0.84,0.91)	(0.62,0.84)
NPV	0.67	0.67	0.83
	(0.55,0.77)	(0.58,0.75)	(0.75,0.89)

Table 7: Estimates of diagnostic accuracy measures.

the bias and improves the performance of the estimators compared to complete case analysis. In addition to this, explicit estimates for the measures of diagnostic accuracy under different missing mechanisms can be computed from simple algebraic formula, while other available methods require iterative methods to estimate the measures of diagnostic accuracy. Furthermore, this approach will allow us to test for MCAR assumption for particular data. However, the only way to confirm for MAR or MNAR missing mechanisms is to recollect the missing data. Although in diagnostic medicine due to the issue of cost-effectiveness, we do not have the luxury of getting hold of the missing data. As a result, careful modeling of missing mechanisms to reduce the bias is more an important issue. Therefore, if the data are not MCAR then by using the scientific knowledge about the data we can assume a MAR or MNAR missing mechanism to compute estimates for the measures of diagnostic accuracy with log-linear models.

Conflict of Interest

The authors have declared no conflict of interest.

References

- Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39: 561-577.
- Nofuentes JAR, Castillo JDLD (2007) The effect of verification bias in the nave estimators of accuracy of a binary diagnostic test. Communications in Statistics: Simulation and Computation 36: 959-972.
- Ransohoff DF, Feinstein AR (1978) Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 299: 926-930.
- Begg CB, Greenes RA (1983) Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 39: 207-215.
- Zhou XH (1998) Comparing accuracies of two screening tests in a two- phase study for dementia. Journal of the Royal Statistical Society: Series C (Applied Statistics) 47: 135-137.
- Rubin DB (1976) Inference and missing data. Biometrika 63: 581-592.
- Zhou XH (1993) Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. Communications in Statistics - Theory and Methods 22: 3177-3198.
- Zhou XH (1994) Effect of verification bias on positive and negative predictive values. Stat Med 13: 1737-1745.
- Harel O, Zhou XH (2006) Multiple imputation for correcting verification bias. Stat Med 25: 3769-3786.
- De Groot JA, Janssen KJ, Zwinderman AH, Bossuyt PM, Reitsma JB, et al. (2011) Correcting for partial verification bias: a comparison of methods. Ann Epidemiol 21: 139-148.
- Baker SG (1995) Evaluating multiple diagnostic tests with partial verification. Biometrics 51: 330-337.
- Kosinski AS, Barnhart HX (2003) Accounting for nonignorable verification bias in assessment of diagnostic tests. Biometrics 59: 163-171.
- Martinez EZ, Achcar JA, Louzada-Neto F (2006) Estimators of sensitivity and specificity in the presence of verification bias: A bayesian approach. Computational Statistics and Data Analysis 51: 601-611.
- Buzoianu M, Kadane JB (2008) Adjusting for verification bias in diagnostic test evaluation: A Bayesian approach. Stat Med 27: 2453-2473.
- Pepe M (2004) The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press.
- Zhou XN, Obuchowski, McClish D (2011) Statistical Methods in Diagnostic Medicine. John Wiley and Sons Inc.
- Baker SG1, Rosenberger WF, Dersimonian R (1992) Closed-form estimates for missing counts in two-way contingency tables. Stat Med 11: 643-657.
- Byrd RH, Lu P, Nocedal J, Zhu C (1995) A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing 16: 1190-1208.
- Drum DE, Christacopoulos JS (1972) Hepatic scintigraphy in clinical decision making. J Nucl Med 13: 908-915.

Appendix

Log-linear parameterization

Expected counts for all verified cases can be modeled by using equation 1 in the following way

$$\begin{aligned} \mu_{ij11} &= \exp\{\lambda + \lambda_i(i) + \lambda_j(j) + \lambda_{R_i}(1) + \lambda_{R_j}(1) + \lambda_{ij}(ij) + \lambda_{iR_i}(i1) + \lambda_{iR_j}(i1) \\ &+ \lambda_{jR_i}(j1) + \lambda_{jR_j}(j1) + \lambda_{R_i R_j}(11)\} \end{aligned}$$

Expected counts for all un-verified cases can be modeled as the following way :

$$\begin{aligned} \mu_{ij12} &= \exp\{\lambda + \lambda_i(i) + \lambda_j(j) + \lambda_{R_i}(1) + \lambda_{R_j}(2) + \lambda_{ij}(ij) + \lambda_{iR_i}(i1) + \lambda_{iR_j}(i2) \\ &+ \lambda_{jR_i}(j1) + \lambda_{jR_j}(j2) + \lambda_{R_i R_j}(12)\} \end{aligned}$$

Some of the side conditions (the sum of each λ -term over each of the indicated subscript is zero) for the log-linear models are

$$\sum_{k=1}^2 \lambda_{R_j}(k) = 0 \quad \sum_{k=1}^2 \lambda_{jR_j}(jk) = 0 \quad \sum_{k=1}^2 \lambda_{iR_j}(jk) = 0 \quad \sum_{k=1}^2 \lambda_{R_i R_j}(1k) = 0$$

By using these conditions, we can get

$$b_{ij} = \frac{\mu_{ij12}}{\mu_{ij11}} = -2 \exp\{\lambda_{R_j}(1) + \lambda_{iR_j}(i1) + \lambda_{jR_j}(j1) + \lambda_{R_i R_j}(11)\}$$

Maximum likelihood (ML) Estimates for Model $(\beta_{..})$ (MCAR)

$$L = \sum_{i,j} n_{ij11} \log(m_{ij}) + \sum_i n_{i+12} \log(m_{i+}\beta_{..}) - \sum_{i,j} m_{ij} (1 + \beta_{..})$$

$$\frac{\partial L}{\partial \beta_{..}} = \sum_i \left(\frac{n_{i+12}}{m_{i+}\beta_{..}} \cdot m_{i+} \right) - \sum_{i,j} m_{ij}$$

$$\therefore 0 = \frac{n_{+12}}{\hat{\beta}_{..}} - \hat{m}_{++} \text{ However, for competed data } \hat{m}_{++} = n_{++11} \text{ which provides the ML estimates for } \beta_{..} \therefore \hat{\beta}_{..} = \frac{n_{+12}}{n_{++11}}$$

ML estimates for m_{ij} for Model $(\beta_{..})$ can be derived by taking the derivative of L with respect to $m_{ij} \therefore \frac{\partial L}{\partial m_{ij}} = \frac{n_{ij11}}{m_{ij}} + \frac{n_{i+12}}{m_{i+}\beta_{..}} \cdot \beta_{..} - (1 + \beta_{..})$

$$\therefore 0 = \frac{n_{ij11}}{\hat{m}_{ij}} + \frac{n_{i+12}}{\hat{m}_{i+}} - (1 + \hat{\beta}_{..})$$

$$\therefore (1 + \hat{\beta}_{..}) = \frac{n_{ij11}}{\hat{m}_{ij}} + \frac{n_{i+12}}{\hat{m}_{i+}}$$

$$\text{Substituting } \hat{\beta}_{..} \text{ from equation ??} \therefore \left(1 + \frac{n_{+12}}{n_{++11}}\right) \cdot \hat{m}_{ij} = n_{ij11} + \frac{n_{i+12}}{\hat{m}_{i+}} \cdot \hat{m}_{ij}$$

$$\therefore \left(\frac{n_{+12}}{n_{++11}}\right) \cdot \hat{m}_{ij} = n_{ij11} + \frac{n_{i+12}}{\hat{m}_{i+}} \cdot \hat{m}_{ij}$$

$$\therefore \hat{m}_{ij} = \left(n_{ij11} + \frac{n_{i+12}}{\hat{m}_{i+}} \cdot \hat{m}_{ij}\right) \cdot \left(\frac{n_{++11}}{n_{++12}}\right) \tag{11}$$

First solving for \hat{m}_{i+}

$$\hat{m}_{i+} = \left(n_{++11} + \frac{n_{i+12}}{\hat{m}_{i+}} \cdot \hat{m}_{i+}\right) \cdot \frac{n_{++11}}{n_{++12}}$$

$$\therefore \hat{m}_{i+} = n_{++11} \cdot \frac{n_{++11}}{n_{++12}}$$

Substituting \hat{m}_{i+} in equation 11

$$\hat{m}_{ij} = \frac{n_{ij11} \cdot n_{++11} \cdot n_{++11}}{n_{++12} \cdot n_{++11}}$$

Maximum likelihood Estimates for Model $(\beta_{i.})$ (MAR)

$$L = \sum_{i,j} n_{ij11} \log(m_{ij}) + \sum_i n_{i+12} \log(m_{i+}\beta_{i.}) - \sum_{i,j} m_{ij} (1 + \beta_{i.})$$

$$\frac{\partial L}{\partial \beta_{i.}} = \frac{n_{i+12}}{m_{i+}\beta_{i.}} \cdot m_{i+} - \sum_j m_{ij}$$

$$\therefore 0 = \frac{n_{i+12}}{\hat{\beta}_i} - \hat{m}_{i+} \text{ Therefore, the MLE estimates for } \hat{\beta}_i \therefore \hat{\beta}_i = \frac{n_{i+12}}{\hat{m}_{i+}}$$

MLE estimates for m_{ij} for Model (β_i) can be derived by taking the derivative of L with respect to $m_{ij} \therefore \frac{\partial L}{\partial m_{ij}} = \frac{n_{ij11}}{m_{ij}} + \frac{n_{i+12}}{m_{i+}\beta_i} \cdot \beta_i - (1 + \beta_i)$

$$\begin{aligned} \therefore 0 &= \frac{n_{ij11}}{\hat{m}_{ij}} + \frac{n_{i+12}}{\hat{m}_{i+}} - (1 + \hat{\beta}_i) \\ \therefore (1 + \hat{\beta}_i) &= \frac{n_{ij11}}{\hat{m}_{ij}} + \frac{n_{i+12}}{\hat{m}_{i+}} \\ \therefore \left(1 + \frac{n_{i+12}}{\hat{m}_{i+}}\right) \cdot \hat{m}_{ij} &= n_{ij11} + \frac{n_{i+12}}{\hat{m}_{i+}} \cdot \hat{m}_{ij} \\ \therefore (\hat{m}_{i+} + n_{i+12}) \cdot \hat{m}_{ij} &= n_{ij11} \hat{m}_{i+} + n_{i+12} \hat{m}_{ij} \\ \therefore \hat{m}_{ij} \cdot (\hat{m}_{i+} + n_{i+12} - n_{i+12}) &= n_{ij11} \hat{m}_{i+} \\ \therefore \hat{m}_{ij} &= n_{ij11} \end{aligned}$$

Maximum likelihood Estimates for $Model(\beta_j)$ (MNAR)

$$\begin{aligned} L &= \sum_{i,j} n_{ij11} \log(m_{ij}) + \sum_i n_{i+12} \log\left(\sum_j m_{ij} \beta_j\right) - \sum_{i,j} m_{ij} (1 + \beta_j) \\ \frac{\partial L}{\partial \beta_j} &= \sum_i \left(\frac{n_{i+12}}{\sum_j m_{ij} \beta_j} \cdot m_{ij} \right) - \sum_i m_{ij} \\ \therefore 0 &= \sum_i \left(\frac{n_{i+12}}{\sum_j \hat{m}_{ij} \hat{\beta}_j} \cdot \hat{m}_{ij} \right) - \sum_i \hat{m}_{ij} \end{aligned}$$

$\therefore \sum_i \hat{m}_{ij} = \sum_i \left(\frac{n_{i+12}}{\sum_j \hat{m}_{ij} \hat{\beta}_j} \cdot \hat{m}_{ij} \right)$ From equation ??, we can guess MLE estimates for $\hat{\beta}_j$ is $\sum_j \hat{m}_{ij} \hat{\beta}_j = n_{i+12} \quad \forall \hat{\beta}_j$. MLE estimates for m_{ij} for Model (β_j) can be derived by taking the derivative of

$$\begin{aligned} \therefore 0 &= \frac{n_{ij11}}{\hat{m}_{ij}} + \frac{n_{i+12}}{\sum_j \hat{m}_{ij} \hat{\beta}_j} \cdot \hat{\beta}_j - (1 + \hat{\beta}_j) \\ \therefore \hat{m}_{ij} (1 + \hat{\beta}_j) &= n_{ij11} + \frac{n_{i+12}}{\sum_j \hat{m}_{ij} \hat{\beta}_j} \cdot \hat{m}_{ij} \hat{\beta}_j \\ \therefore \hat{m}_{ij} &= n_{ij11} \text{ We can easily check our guess satisfies all the MLE equation for Model } (\beta_j) \end{aligned} \tag{12}$$

Comparison of MAR Estimates with Zhou's Result

The \square estimates for Sensitivity is

$$\begin{aligned} Sensitivity(Zhou) &= \frac{\frac{(n_{1+} + n_{111})}{n_{+1+}(n_{111} + n_{121})}}{\frac{n_{1+} n_{111}}{n_{+1+}(n_{111} + n_{121})} + \frac{n_{2+} n_{211}}{n_{+1+}(n_{211} + n_{221})}} \\ &= \frac{(n_{1+} + n_{111}) n_{2+1}}{n_{1+} n_{111} n_{2+1} + n_{2+} n_{211} n_{1+1}} \end{aligned} \tag{13}$$

From equation 6 the Sensitivity estimates from log-linear model are

$$Sensitivity(MAR) = \frac{\hat{m}_{11} (1 + \hat{\beta}_1)}{\hat{m}_{11} (1 + \hat{\beta}_1) + \hat{m}_{21} (1 + \hat{\beta}_2)}$$

$$\begin{aligned}
&= \frac{n_{1111} \left(1 + \frac{n_{1+12}}{\hat{m}_{1+}} \right)}{n_{1111} \left(1 + \frac{n_{1+12}}{\hat{m}_{1+}} \right) + n_{2111} \left(1 + \frac{n_{2+12}}{\hat{m}_{2+}} \right)} \\
&= \frac{n_{1111} \left(1 + \frac{n_{1+12}}{n_{1+11}} \right)}{n_{1111} \left(1 + \frac{n_{1+12}}{n_{1+11}} \right) + n_{2111} \left(1 + \frac{n_{2+12}}{n_{2+11}} \right)} \\
&= \frac{(n_{1+1+} n_{1111}) n_{2+11}}{n_{1+1+} n_{1111} n_{2+11} + n_{2+1+} n_{2111} n_{1+11}} = \text{Sensitivity(Zhou)}
\end{aligned}$$

Similarly, we can show that the estimates for Specificity from log-linear models are equivalent to [] Specificity estimates.

Information Matrices :

Model ($\beta_{..}$)

$$\begin{aligned}
\frac{\partial^2 L}{\partial \beta_{..}^2} &= -\frac{n_{++12}}{\beta_{..}^2} \\
-E \left(\frac{\partial^2 L}{\partial \beta_{..}^2} \right) &= E \left(\frac{n_{++12}}{\beta_{..}^2} \right) = \frac{m_{++}\beta_{..}}{\beta_{..}^2} = \frac{m_{++}}{\beta_{..}} \\
-E \left(\frac{\partial^2 L}{\partial m_{ij}} \right) &= \frac{1}{m_{ij}} + \frac{\beta_{..}}{m_{i+}} \\
-E \left(\frac{\partial^2 L}{\partial m_{ij}\beta_{..}} \right) &= 1
\end{aligned}$$

$$\mathbf{I}_1 = \begin{pmatrix} -E \left(\frac{\partial^2 L}{\partial m_{11}^2} \right) & -E \left(\frac{\partial^2 L}{\partial m_{11}m_{12}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{11}m_{21}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{11}m_{22}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{11}\beta_{..}} \right) \\ -E \left(\frac{\partial^2 L}{\partial m_{12}m_{11}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{12}^2} \right) & -E \left(\frac{\partial^2 L}{\partial m_{12}m_{21}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{12}m_{22}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{12}\beta_{..}} \right) \\ -E \left(\frac{\partial^2 L}{\partial m_{21}m_{11}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{21}m_{12}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{21}^2} \right) & -E \left(\frac{\partial^2 L}{\partial m_{21}m_{22}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{21}\beta_{..}} \right) \\ -E \left(\frac{\partial^2 L}{\partial m_{22}m_{11}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{22}m_{12}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{22}m_{21}} \right) & -E \left(\frac{\partial^2 L}{\partial m_{22}^2} \right) & -E \left(\frac{\partial^2 L}{\partial m_{22}\beta_{..}} \right) \\ -E \left(\frac{\partial^2 L}{\partial \beta_{..}m_{11}} \right) & -E \left(\frac{\partial^2 L}{\partial \beta_{..}m_{12}} \right) & -E \left(\frac{\partial^2 L}{\partial \beta_{..}m_{21}} \right) & -E \left(\frac{\partial^2 L}{\partial \beta_{..}m_{22}} \right) & -E \left(\frac{\partial^2 L}{\partial \beta_{..}^2} \right) \end{pmatrix}$$

Model ($\beta_{i.}$)

$$\begin{aligned}
-E \left(\frac{\partial^2 L}{\partial \beta_{i.}^2} \right) &= \frac{m_{i+}}{\beta_{i.}} \\
-E \left(\frac{\partial^2 L}{\partial m_{ij}} \right) &= \frac{1}{m_{ij}} + \frac{\beta_{i.}}{m_{i+}} \\
-E \left(\frac{\partial^2 L}{\partial m_{ij}\beta_{i.}} \right) &= 1
\end{aligned}$$

