Published in final edited form as:
Comput Stat Data Anal. 2008 September ; 53(1): 27–37. doi:10.1016/j.csda.2008.05.031.

Comparing Multiple Sensitivities and Specificities with Different Diagnostic Criteria: Applications to Sexual Abuse and Sexual Health Research

Q. Yu[1], **W. Tang**[1,2], **Y. Ma**[1], **S.A. Gamble**[2], and **X.M. Tu**[1,2]

[1]Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642

[2]Department of Psychiatry, University of Rochester, Rochester, NY 14642

Summary

When comparing sensitivities and specificities from multiple diagnostic tests, particularly in biomedical research, the different test kits under study are applied to groups of subjects with the same disease status for a disease or medical condition under consideration. Although this process gives rise to clustered or correlated test outcomes, the associated inference issues are well recognized and have been widely discussed in the literature. In mental health and psychosocial research, sensitivity and specificity have also been widely used to study the reliability of instrument for diagnosing mental health and psychiatric conditions and assessing certain behavioral patterns. However, unlike biomedical applications, outcomes are often obtained under varying reference standards or different diagnostic criteria, precluding the application of existing methods for comparing multiple diagnostic tests to such a research setting. In this paper, we develop a new approach to address these problems (including that of missing data) by extending recent work on inference using inverse probability weighted estimates. The approach is illustrated with data from two studies in sexual abuse and health research as well as a limited simulation study, with the latter used to study the performance of the proposed procedure.

Keywords

Diagnostic test; Inverse probability weighted estimate; Missing data; Positive and negative predictive value; Psychosocial research

1 Introduction

Diagnostic tests are widely used in biomedical research to detect certain medical conditions or diseases in populations of interest. The accuracy of a diagnostic test is defined by comparing the test result to the true condition or disease status of the subject tested. The most commonly used measures in evaluating the accuracy of diagnostic test are test sensitivity and specificity. In biomedical research, such measures are used to evaluate and compare the quality of different test kits that are designed to detect a common disease of interest such as HIV (e.g. Cross et al.

1992; Johnson and Gastwirth, 1991; Kowalski et al. 2001; Taylor et al. 1990; Tu et al. 1992; 1995). In most studies, different diagnostic tests are applied to a group of diseased or non-diseased subjects for such evaluations (Ahn, 1997, Kowalski et al. 2001; Lee and Dubin, 1994; Mendoza-Blanco et al. 1996; Yasemin, 2005). Since the multiple outcomes from the different tests are based on the same individual, they form clustered or correlated responses. Inferences for such correlated multivariate binary responses have been discussed in the literature (Ahn, 1997, Genc et al. 2005; Kowalski et al. 2001; Lee and Dubin, 1994; Mendoza-Blanco et al. 1996; Taylor et al. 1990; Yasemin, 2005).

In addition to biomedical applications, sensitivity and specificity have also been widely used in mental health and psychosocial research to study the reliability of instruments for diagnosing mental health and psychiatric conditions and for measuring behavioral patterns and past health-related histories. In such applications, the need to evaluate and compare different instruments arises. However, unlike biomedical applications, these multiple instruments may involve varying reference standards or diagnostic criteria, precluding the application of existing methods for comparing multiple diagnostic tests to such a research setting. For example, in sexual abuse research, identifying whether others have knowledge of a patient's childhood sexual abuse history is important from both a methodological and clinical perspective (Gamble et al., in press; Lipschitz et al., 1999; Talbot et al. 2004). In such studies, an important issue is reliability of the information about the proband's sexual abuse history provided by a family member or informant. Research shows that subjects and informants have highly concordant reports about severe childhood sexual abuse, but not about less severe or more infrequent sexual abuse experiences (Gamble et al., 2006). Because the subject's response (reference standard) about her/his abuse experience varies depending on how sexual abuse is defined (i.e., more severe or less severe), inference procedures developed for biomedical applications with a static or single reference standard no longer apply.

In this paper, we describe a new approach to address the inference problems inherent in formal comparison of the sensitivity (specificity) estimates derived based on multiple diagnostic criteria such as severity of sexual abuse. In particular, we discuss how to address the impact of missing data by extending recent work on inference using inverse probability weighted estimates. We illustrate the methodology with both real and simulated study data.

2 Models for Comparing Multiple Sensitivities and Specificities

In this section, we first briefly review the statistical issues that arise when comparing multiple diagnostic tests in biomedical applications. We then discuss how the issues are different when comparing multiple sensitivities and specificities defined by different diagnostic criteria and develop new approaches to address them.

2.1 Multiple Tests under a Common Reference Standard

In most applications in biomedical research, different diagnostic tests are applied to a group of subjects with the same disease status. For example, to compare sensitivities, the tests under study are applied to a group of subjects with the disease. Likewise, to compare specificities, different tests are administered to a group of disease-free subjects.

Consider m diagnostic tests. Let D (D^c) denote the disease (non-disease) status and $T_k^+ \left(T_k^-\right)$ denote the positive (negative) test by the kth test kit ($1 \leq k \leq m$). In most biomedical studies, each test kit is applied to a sample of diseased (non-diseased) subjects to derive data for estimating and comparing sensitivities (specificities) across the test kits. We focus on sensitivity, since the consideration for specificity is similar.

Let y_{ki} be a binary variable denoting the outcome of the kth test kit when applied to the ith subject from the diseased group, with the value 1 (0) denoting a positive (negative) test outcome. For each test kit, sensitivity is defined as the probability of a positive test given the disease

$$\phi_k=\Pr\left[T_k^+|D\right]=\Pr\left[y_{ki}=1|D\right]=E\left(y_{ki}|D\right),\ 1\le k\le m. \tag{1}$$

With test data from the diseased group, $\boldsymbol{\phi}_k$ is readily estimated by:

$\widehat{\phi}_k=\frac{r_k}{n}=\frac{1}{n}\sum_{i=1}^{n}y_{ki}\ (1\le k\le m)$, where r_k is the number of positive tests by the kth kit. Now let

$$\mathbf{y}_i=(y_{1i},y_{2i}\ldots,y_{mi})^\top,\ \theta=(\phi_1,\phi_2,\ldots,\phi_m)^\top,\ 1\le i\le n,$$

By applying (1) and the estimate for each test kit to $\boldsymbol{\theta}$, we obtain an estimate of the sensitivity vector: $\widehat{\theta}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$. The asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is readily obtained by invoking the central limit theorem (CLT)

$$\sqrt{n}\left(\widehat{\theta}-\theta\right)\to_d N\left(0,\Sigma_\theta=E\left[(\mathbf{y}_i-\theta)(\mathbf{y}_i-\theta)^\top\right]\right),\quad n\to\infty, \tag{2}$$

where $\to_{\mathrm{d}}$ denotes convergence in distribution (e.g., Kowalski and Tu, Chap. 1, 2007). By applying CLT and Slutsky's theorem (e.g., Kowalski and Tu, Chap. 1, 2007), we obtain a consistent estimate of the asymptotic variance Σ_θ, $\widehat{\Sigma}_\theta=\frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\mathbf{y}_i-\widehat{\theta}\right)\left(\mathbf{y}_i-\widehat{\theta}\right)^\top\right]$.

For any (smooth) function $\mathbf{g}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ (i.e., $\mathbf{g}(\boldsymbol{\theta})$ has continuous first-order derivatives), by the Delta method (e.g., Kowalski and Tu, Chap. 1, 2007), the statistic $g(\hat{\boldsymbol{\theta}})$ has the following distribution:

$$\sqrt{n}\left[\mathbf{g}\left(\widehat{\theta}\right)-\mathbf{g}(\theta)\right]\to_d N\left(0,\Sigma_{g(\theta)}=\frac{\partial}{\partial\theta}\mathbf{g}(\theta)\Sigma_\theta\frac{\partial^\top}{\partial\theta}\mathbf{g}(\theta)\right), \tag{3}$$

where $\frac{\partial}{\partial\theta}\mathbf{g}(\theta)$ denotes the derivative of $\mathbf{g}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and $\frac{\partial^\top}{\partial\theta}\mathbf{g}(\theta)$ the transpose of $\frac{\partial}{\partial\theta}\mathbf{g}(\theta)$. By estimating $\Sigma_{g(\theta)}$ using a consistent estimate such as $\widehat{\Sigma}_{g(\theta)}=\frac{\partial}{\partial\theta}\mathbf{g}\left(\widehat{\theta}\right)\widehat{\Sigma}_\theta\frac{\partial^\top}{\partial\theta}\mathbf{g}\left(\widehat{\theta}\right)$, we can use (3) to make inference about $\mathbf{g}(\boldsymbol{\theta})$. For example, many hypotheses of practical interest concerning $\boldsymbol{\theta}$ can be expressed in terms of a linear contrast:

$$H_0:K\theta=\mathbf{0},\quad \text{vs.}\quad H_a:K\theta\ne\mathbf{0}, \tag{4}$$

where K is some $l\times m$ full rank matrix with known constants ($l\le m$). By setting $\mathbf{g}(\boldsymbol{\theta})=K\boldsymbol{\theta}$ and applying (3), we can immediate obtain the asymptotic distribution of $K\hat{\boldsymbol{\theta}}$ for testing **(4)**. In particular, under H_0, the quadratic statistic, $Q_n^2=n\widehat{\theta}^\top K^\top\left(K\widehat{\Sigma}_\theta K^\top\right)^{-1}K\widehat{\theta}$, has an asymptotic central χ^2 distribution with l degrees of freedom. For example, the null of no difference across three test kits, i.e., $H_0:\phi_1=\phi_2=\phi_3$, can be expressed as a linear contrast with

$K=\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$. It follows that $Q_n^2=n\widehat{\theta}^\top K^\top \left(K\widehat{\Sigma}_\theta K^\top\right)^{-1} K\widehat{\theta}$ has an asymptotic central χ^2 distribution with 2 degrees of freedom.

In addition to test sensitivity (specificity), the positive (PPV) and negative (NPV) predictive values are also widely used in biomedical and epidemiologic studies to indicate the degree of accuracy when ascertaining disease status based on test outcomes (Fleiss et al. 2003; Tu et al. 1992). PPV (NPV) is the probability of disease (non-disease) given a positive (negative) test. As both PPV and NPV are a function of disease prevalence (in addition to sensitivity and specificity), their values reflect not only the test kit accuracy, but also the disease prevalence as well. In most biomedical applications, PPV (NPV) is not directly estimated from data, as samples selected for evaluating sensitivity (specificity) typically contain diseased (disease-free) subjects rather than a random sample from a population of interest (Tu et al. 1992; Tu et al. 1994; Fleiss et al. 2003). This likely explains the paucity of literature on inference for PPV (NPV). In contrast, samples selected for evaluating diagnostic tests in behavioral and psychosocial research are often random samples (or can be interpreted as such) from the study population of interest and as a result, it is of interest to estimate PPV (NPV).

Now, consider comparing multiple PPVs based on test outcomes from a random sample consisting of both diseased and disease-free subjects. Let n_k denote the number of positive tests and r_k the number of diseased subjects among those who test positive by the kth test kit ($1 \leq k \leq m$). Then, $\widehat{\omega}_k=\frac{r_k}{n_k}$ is a consistent estimate of PPV for the kth test kit, $\omega_k=\Pr[D|T_k^+]$, with the following asymptotic distribution:

$$\sqrt{n_k}\,(\widehat{\omega}_k-\omega_k)\rightarrow_d N\left(0,\sigma_k^2=\omega_k\,(1-\omega_k)\right),\quad k=1,2,\ldots,m. \tag{5}$$

The above can be used for inference about each individual ω_k. To compare PPVs across the different test kits, we need the joint asymptotic distribution of the vector statistic, $\hat{\omega} = (\hat{\omega}_1,\ldots,\hat{\omega}_m)^\top$. Although similar in form, this joint distribution requires quite different considerations from those for sensitivity. Interestingly, these are exactly the same problems that arise when comparing multiple sensitivities (specificities) with varying diagnostic criteria, which we discuss next.

2.2 Multiple Tests with Different Diagnostic Criteria

To illustrate the underlying issues, consider again the sexual abuse study in the Introduction. Let n_1 denote the number of subjects who answered yes and r_1 the number of informants who corroborated the subjects' responses to the question whether the subject had any sexual abuse (including severe, moderately severe, and less severe sexual abuse). By treating the subject's response as a gold standard, the sensitivity of informant's response is estimated by $\widehat{\psi}_1=\frac{r_1}{n_1}$. Now, consider the question of whether the subject had severe sexual abuse. Let n_2 denote the number of subjects who answered yes and r_2 the number of informants who concurred with the responses. Again using the subject's response as a gold standard, the sensitivity of the informant's response is estimated by $\widehat{\psi}_2=\frac{r_2}{n_2}$. By comparing with $\widehat{\phi}_k=\frac{r_k}{n}$ discussed in the preceding section, it is seen that $\hat{\psi}_k$ have a varying denominator n_k between the two questions (or diagnostic criteria). As in the case of PPV, $\hat{\boldsymbol{\theta}} = (\hat{\psi}_1,\hat{\psi}_2)^\top$ depends not only on r_k, but on n_k as well.

To develop the joint asymptotic distribution of $\hat{\boldsymbol{\theta}}$, let z_{ki} and y_{ki} be a binary variable denoting the response (1 for yes and 0 for no) from the ith subject and informant pair for the k's th question (k = 1; 2). Then, the sensitivity of the kth question is given by

$$\psi_k=\Pr\left[y_{ki}=1 \,\middle|\, z_{ki}=1\right]=\frac{\Pr\left[y_{ki}=1,\ z_{ki}=1\right]}{\Pr\left[z_{ki}=1\right]}=\frac{E\left(y_{ki}z_{ki}\right)}{E\left(z_{ki}\right)},\quad k=1,2. \tag{6}$$

In comparison to (1), the sensitivity ψ_k has a more complex expression; it depends not only on the test outcome (y_{ki}), but on the gold standard (z_{ki}) as well. By substituting the moment estimates $\frac{1}{n}\sum_{i=1}^{n} y_{ki}z_{ki}$ and $\frac{1}{n}\sum_{i=1}^{n} z_{ki}$ in place of the respective means, we immediately obtain the estimates $\widehat{\psi}_k=\frac{\sum_{i=1}^{n} y_{ki}z_{ki}}{\sum_{i=1}^{n} z_{ki}}$.

To derive their joint asymptotic distribution, let

$$\begin{aligned}
\mathbf{w}_{ki} &= (z_{ki}, y_{ki}z_{ki})^{\top},\ \boldsymbol{\zeta}_k=E\left(\mathbf{w}_{ki}\right),\ \widehat{\boldsymbol{\zeta}}_k=\tfrac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{ki},\ \mathbf{w}_i=\left(\mathbf{w}_{1i}^{\top}, \mathbf{w}_{2i}^{\top}\right)^{\top},\\
\boldsymbol{\zeta} &= \left(\boldsymbol{\zeta}_1^{\top}, \boldsymbol{\zeta}_2^{\top}\right)^{\top},\ \widehat{\boldsymbol{\zeta}}=\left(\widehat{\boldsymbol{\zeta}}_1^{\top}, \widehat{\boldsymbol{\zeta}}_2^{\top}\right)^{\top}=\tfrac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i,\ \psi_k=f_k(\boldsymbol{\zeta}_k)=\tfrac{\zeta_{k2}}{\zeta_{k1}},\\
\theta &= (\psi_1,\psi_2)^{\top}=(f_1(\boldsymbol{\zeta}_1), f_2(\boldsymbol{\zeta}_2))^{\top}=\mathbf{f}(\boldsymbol{\zeta}).
\end{aligned} \tag{7}$$

Then, we can express $\hat{\boldsymbol{\theta}}$ as a function of $\hat{\boldsymbol{\zeta}}$

$$\widehat{\theta}=\left(\widehat{\psi}_1,\widehat{\psi}_2\right)^{\top}=f\left(\widehat{\boldsymbol{\zeta}}\right),\quad \widehat{\psi}_k=f_k\left(\widehat{\boldsymbol{\zeta}}_k\right)=\frac{\widehat{\zeta}_{k2}}{\widehat{\zeta}_{k1}}=\frac{\frac{1}{n}\sum_{i=1}^{n} y_{ki}z_{ki}}{\frac{1}{n}\sum_{i=1}^{n} z_{ki}}. \tag{8}$$

By applying CLT to $\hat{\boldsymbol{\zeta}}$ and the Delta method to $f(\hat{\boldsymbol{\zeta}})$, we obtain the asymptotic distribution of $\hat{\boldsymbol{\theta}}$

$$\sqrt{n}\left(\widehat{\theta}-\theta\right)\to_d N\left(0,\Sigma_{\theta}=\frac{\partial\mathbf{f}}{\partial\boldsymbol{\zeta}}\Sigma_{\zeta}\frac{\partial^{\top}\mathbf{f}}{\partial\boldsymbol{\zeta}}\right),\quad \Sigma_{\zeta}=\mathrm{Var}\left(\mathbf{w}_i\right),\quad n\to\infty. \tag{9}$$

A consistent estimate of the asymptotic variance Σ_θ is given by:

$$\widehat{\Sigma}_{\theta}=\frac{\partial}{\partial\boldsymbol{\zeta}}f\left(\widehat{\boldsymbol{\zeta}}\right)\widehat{\Sigma}_{\zeta}\frac{\partial^{\top}}{\partial\boldsymbol{\zeta}}f\left(\widehat{\boldsymbol{\zeta}}\right),\quad \widehat{\Sigma}_{\zeta}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{w}_i-\widehat{\boldsymbol{\zeta}}\right)\left(\mathbf{w}_i-\widehat{\boldsymbol{\zeta}}\right)^{\top}. \tag{10}$$

We can readily extend the above development to a general setting with more than two diagnostic criteria. Consider m diagnostic criteria and let z_{ki} (y_{ki}) be defined as above except that k now ranges from 1 to m. Also, let $\mathbf{w}_{ki}$, $\mathbf{w}_i$, $\boldsymbol{\zeta}_k$, $\hat{\boldsymbol{\zeta}}_k$, $\boldsymbol{\zeta}$, $\hat{\boldsymbol{\zeta}}$, ψ_k, $\hat{\psi}_k$, $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ and $f(\boldsymbol{\zeta})$ be defined as in (7), but with k ranging from 1 to m. By applying **(9)** and **(10)** to $\hat{\boldsymbol{\theta}}$, we obtain the distribution of $\hat{\boldsymbol{\theta}}$ and a consistent estimate of the asymptotic variance. As in Section 2.1, we can test any linear contrast involving $\boldsymbol{\theta}$ such as equal sensitivities across different diagnostic criteria based

on the distribution of $\hat{\boldsymbol{\theta}}$. We again use the quadratic statistic, $Q_n^2 = n\widehat{\theta}^\top K^\top \left(K\widehat{\Sigma}_\theta K^\top\right)^{-1} K\widehat{\theta}$, and the associated χ^2 distribution for inference.

By using a similar argument, we can find the joint asymptotic distribution of $\hat{\boldsymbol{\omega}}$ for comparing multiple PPVs as discussed in Section 2.1. Since the PPV for the kth criterion is

$$\omega_k = \Pr[z_{ki}=1 \mid y_{ki}=1] = \frac{E(y_{ki}z_{ki})}{E(y_{ki})}, \quad 1 \le k \le m,$$

we readily obtain the asymptotic distribution of $\hat{\boldsymbol{\omega}}$ and a consistent estimate of the asymptotic variance by reversing the roles of y_{ki} and z_{ki} in the above development. The discussion for specificity and NPV is similar.

2.3 Missing Data

Missing data are a common problem in research. In this section, we discuss the impact of missing data and how to ensure valid inference in its presence. Again, for convenience, we focus on test sensitivity.

Within the current context, the kth sensitivity ψ_k in (6) is defined by the parameter vector $\boldsymbol{\zeta}_k = (E(z_{ki}), E(y_{ki}z_{ki}))^\top$. If the component $z_{ki}, y_{ki}z_{ki}$ or both are inconsistently estimated, the resulting estimate $\hat{\psi}_k$ is generally biased. To ensure valid inference, we must construct consistent estimates of $\boldsymbol{\zeta}_k$.

In general, missing data may occur to z_{ki}, y_{ki} or both. For estimating $E(z_{ki})$, we must have non-missing z_{ki}, while for $E(y_{ki}z_{ki})$, we must have both y_{ki} and z_{ki} observed. To help construct and discuss estimates, we define a set of indicators for missing (or rather observed) data as follows

$$\begin{aligned}
r_{kyi} &= \begin{cases} 1 & \text{if } y_{ki} \text{ is observed} \\ 0 & \text{if otherwise} \end{cases}, \quad r_{kzi} = \begin{cases} 1 & \text{if } z_{ki} \text{ is observed} \\ 0 & \text{if otherwise} \end{cases}, \\
R_{ki} &= \begin{pmatrix} r_{kzi} & 0 \\ 0 & r_{kyi}r_{kzi} \end{pmatrix}, \quad R_i = \operatorname{diag}(R_{ki}) = \begin{pmatrix} R_{1i} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & R_{mi} \end{pmatrix}, \\
\mathbf{w}_{ki} &= \begin{cases} (0,0)^\top & \text{if } z_{ki} \text{ or both } z_{ki} \text{ and } y_{ki} \text{ are missing} \\ (z_{ki},0)^\top & \text{if only } y_{ki} \text{ is missing} \\ (z_{ki}, y_{ki}z_{ki})^\top & \text{if otherwise} \end{cases}, \quad \mathbf{w}_i = \left(\mathbf{w}_{1i}^\top, \cdots, \mathbf{w}_{mi}^\top\right)^\top.
\end{aligned} \tag{11}$$

Note that in many studies, z_{ki} (or y_{ki}) are either observed or missing together for all k ($1 \le k \le m$). For example, in the sexual abuse study, z_{ki} (y_{ki}) were obtained based on the ith proband's (informant's) response under some cut-point on severity of sexual abuse (see also Example 1 in Section 3 for more details) and the occurrence of missing data do not depend on the different diagnostic criteria. Throughout the rest of discussion, we assume $r_{kzi} = r_{zi}$ ($r_{kyi} = r_{yi}$) for all k ($1 \le k \le m$).

One way to estimate $\boldsymbol{\zeta}_k$ is to compute the sample means of $E(z_{ki})$ and $E(y_{ki}z_{ki})$ based on available data. Using the missing data indicators, such an estimate can be expressed elegantly as $\widehat{\boldsymbol{\zeta}} = \left(\sum_{i=1}^n R_i\right)^{-1} \sum_{i=1}^n R_i \mathbf{w}_i$. This estimate is consistent if missing data follow the missing completely at random (MCAR) assumption (Rubin, 1976). Since under MCAR r_{yi} and r_{zi} are independent of z_{ki} and y_{ki}, it follows from the law of large numbers (LLN) and Slutsky's theorem that

$$\widehat{\boldsymbol{\zeta}} \to_p E^{-1}(R_i)\, E\,(R_i \mathbf{w}_i) = E^{-1}(R_i)\, E\,(R_i)\, E\,(\mathbf{w}_i) = \boldsymbol{\zeta},$$

where $\to_p$ denotes convergence in probability (e.g., Kowalski and Tu, Chap. 1, 2007).

However, when MCAR fails, i.e., if r_{yi}, r_{zi} or both depend on z_{ki}, y_{ki} and some covariates, $\hat{\boldsymbol{\zeta}}$ above is likely to be inconsistent. For example, in the sexual abuse study, if an informant indicated that he/she did not know whether the subject had any sexual abuse, such a missing response might reflect the fact that the informant did not know the subject well enough to have such knowledge, giving rise to the dependence of r_{yi} on covariates that measure the relationship between the proband and informant. In this case, estimating $E\,(y_{ki} z_{ki})$ based simply on the observed z_{ki} and y_{ki} may yield biased estimates.

To obtain consistent estimates in such scenarios, let

$$\begin{aligned}
\mathbf{y}_i &= (y_{1i}, \ldots, y_{mi})^\top, \quad \mathbf{z}_i = (z_{1i}, \ldots, z_{mi})^\top, \\
\pi_{zi} &= \Pr\,[\, r_{zi} = 1 \mid \mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i], \quad \pi_{zyi} = \Pr\,[\, r_{zi} = 1, r_{yi} = 1 \mid \mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i], \\
D_{ki} &= \begin{pmatrix} \pi_{zi}^{-1} r_{zi} & 0 \\ 0 & \pi_{zyi}^{-1} r_{yi} r_{zi} \end{pmatrix}, \quad D_i = \begin{pmatrix} D_{1i} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & D_{mi} \end{pmatrix},
\end{aligned}$$

where $\mathbf{x}_i$ is a $q \times 1$ column vector of covariates with no missing data. Consider a revised inverse probability weighted (IPW) estimate of $\boldsymbol{\zeta}$ as follows: $\widehat{\boldsymbol{\zeta}} = \frac{1}{n} \sum_{i=1}^{n} D_i \mathbf{w}_i$. (Note that for notational brevity, we still used $\hat{\boldsymbol{\zeta}}$ to denote the resulting estimate.) It is readily checked that

$$\begin{aligned}
E\left(\frac{r_{zi}}{\pi_{zi}} z_{ki}\right) &= E\left[\frac{z_{ki}}{\pi_{zi}} E\,(r_{zi} \mid \mathbf{y}_i, \mathbf{z}_i)\right] = E\,(z_{ki}), \quad E\,(D_{ki}) = \mathbf{I}_2, \\
E\left(\frac{r_{zi} r_{yi}}{\pi_{zyi}} y_{ki} z_{ki}\right) &= E\left[\frac{y_{ki} z_{ki}}{\pi_{zyi}} E\,(r_{zi} r_{yi} \mid \mathbf{y}_i, \mathbf{z}_i)\right] = E\,(y \qquad\qquad {}_{ki} z_{ki}),
\end{aligned} \tag{12}$$

where $\mathbf{I}_2$ denotes the 2×2 identity matrix. It then follows from (12) that

$$\widehat{\boldsymbol{\zeta}} \to_p E\,(D_i \mathbf{w}_i) = \mathrm{diag}_k\,(E\,(D_{ki} \mathbf{w}_{ki})) = \mathrm{diag}_k \left(\begin{pmatrix} E\left(\frac{r_{zi}}{\pi_{zi}} z_{ki}\right) & 0 \\ 0 & E\left(\frac{r_{zi} r_{yi}}{\pi_{zyi}} y_{ki} z_{ki}\right) \end{pmatrix} \right) = \boldsymbol{\zeta},$$

where $diag_k\,(A_k)$ denotes a block-diagonal matrix with A_k on the kth diagonal. Thus, the revised estimate $\hat{\boldsymbol{\zeta}}$ is consistent. By CLT, we obtain the asymptotic distribution and a consistent estimate of the asymptotic variance

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\right) &\to_d N\,(0, \Sigma_\zeta = V\,ar\,(D_i \mathbf{w}_i)), \\
\widehat{\Sigma}_\zeta &= \tfrac{1}{n-1} \left[\sum_{i=1}^{n} \left(D_i \mathbf{w}_i - \widehat{\boldsymbol{\zeta}}\right) \left(D_i \mathbf{w}_i - \widehat{\boldsymbol{\zeta}}\right)^\top \right].
\end{aligned} \tag{13}$$

As in Section 2.2, we can construct a consistent estimate of $\boldsymbol{\theta}$ based on $\hat{\boldsymbol{\zeta}}$ as well as a consistent estimate of the asymptotic variance of $\hat{\boldsymbol{\theta}}$ by applying the Delta method to (13).

If π_{zi} and π_{zyi} are known, $\hat{\boldsymbol{\zeta}}$ and $\hat{\Sigma}_{\zeta}$ are readily computed. However, in most studies, these quantities are unknown and must be estimated. By viewing the event $\{r_{zi} = 1; r_{yi} = 1\}$ as one of the possible outcomes below

$$\begin{aligned} \omega_{i1} &= \{r_{zi}=1, r_{yi}=1\}, \quad \omega_{i2} = \{r_{zi}=0, r_{yi}=1\}, \\ \omega_{i3} &= \{r_{zi}=1, r_{yi}=0\}, \quad \omega_{i4} = \{r_{zi}=0, r_{yi}=0\}, \end{aligned}$$

we can estimate π_{zi} and π_{zyi} by modeling the categorical outcomes ω_{ij} above using the generalized logit model (e.g., Kowalski and Tu, Chap. 2, 2007).

Let $I_{ij} = 1$ if ω_{ij} is observed and 0 if otherwise. The generalized logit model has the following form

$$\begin{aligned} p_{ij}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i; \eta) &= E(I_{ij}=1 \mid \mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i) = \frac{\exp\left(\eta_{oj}+\eta_{xj}^{\top}\mathbf{x}_i+\eta_{zj}^{\top}\mathbf{z}_i+\eta_{yj}^{\top}\mathbf{y}_i\right)}{\Phi(\mathbf{x}_i,\mathbf{z}_i\mathbf{y}_i;\eta)}, \\ \eta_{xj} &= \left(\eta_{xj1}, \ldots, \eta_{xjq}\right)^{\top}, \quad \eta_{zj} = \left(\eta_{zj1}, \ldots, \eta_{zjm}\right)^{\top}, \quad \eta_{yj} = \left(\eta_{yj1}, \ldots, \eta_{yjm}\right)^{\top}, \\ \eta_{01} &= 0, \quad \eta_{x1} = \mathbf{0}, \quad \eta_{zj} = \eta_{yj} = \mathbf{0}, \quad 1 \le j \le 4, \\ \Phi(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i; \eta) &= \sum_{j=1}^{4} \exp\left(\eta_j+\eta_{xj}^{\top}\mathbf{x}_i+\eta_{zj}^{\top}\mathbf{z}_i+\eta_{yj}^{\top}\mathbf{y}_i\right), \quad \eta_x = \left(\eta_{x2}^{\top}, \eta_{x3}^{\top}, \eta_{x4}^{\top}\right)^{\top}, \\ \eta_z &= \left(\eta_{z2}^{\top}, \eta_{z3}^{\top}, \eta_{z4}^{\top}\right)^{\top}, \quad \eta_y = \left(\eta_{y2}^{\top}, \eta_{y3}^{\top}, \eta_{y4}^{\top}\right)^{\top}, \quad \eta = \left(\eta_{02}, \eta_{03}, \eta_{04}, \eta_x^{\top}, \eta_z^{\top}, \eta_y^{\top}\right)^{\top}. \end{aligned} \tag{14}$$

However, as shown in the Appendix, it is generally not possible to model p_{ij} as a function of $\mathbf{z}_i$ and $\mathbf{y}_i$ in addition to $\mathbf{x}_i$. In other words, $p_{ij}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i; \eta)$ can be a function of $\mathbf{x}_i$ only, i.e., $p_{ij}(\mathbf{x}_i; \mathbf{y}_i; \mathbf{z}_i; \eta) = p_{ij}(\mathbf{x}_i; \eta)$, in which case (14) reduces to

$$\begin{aligned} p_{ij}(\mathbf{x}_i;\eta) &= E(I_{ij}=1 \mid \mathbf{x}_i) = \frac{\exp\left(\eta_{oj}+\eta_{xj}^{\top}\mathbf{x}i\right)}{\Phi(\mathbf{x}_i;\eta)}, \\ \eta_{xi} &= (\eta_{xj1}, \ldots, \eta_{xjq})^{\top}, \quad \eta_{01} = \mathbf{0}, \quad \eta_{x1} = \mathbf{0}, \quad 1 \le j \le 4, \\ \Phi(\mathbf{x}_i;\eta) &= \sum_{j=1}^{4} \exp\left(\eta_j+\eta_{xj}^{\top}\mathbf{x}_i\right), \quad \eta = \left(\eta_{02}, \eta_{03}, \eta_{04}, \eta_x^{\top}\right)^{\top}. \end{aligned} \tag{15}$$

Under (15), $\pi_{zyi} = p_{i1}(\mathbf{x}_i; \boldsymbol{\eta})$ and $\pi_{zi} = p_{i1}(\mathbf{x}_i; \boldsymbol{\eta}) + p_{i3}(\mathbf{x}_i; \boldsymbol{\eta})$. Thus, we can first estimate $\boldsymbol{\eta}$ using the maximum likelihood procedure and then estimate $p_{ij}(\mathbf{x}_i; \boldsymbol{\eta})$ by substituting such estimates in place of $\boldsymbol{\eta}$.

Note that as a special case, if the gold standard $\mathbf{z}_i$ is observed for all subjects, inference may be facilitated by applying the generalized linear mixed effects model (GLMM) or the generalized estimating equations (GEE) with a logit link (e.g., Kowalski et al., 2001; Leisenring et al. 2000). Note also that a model similar to (14) has been used to model non-ignorable (non-MAR) missingness for bivariate binary responses under likelihood based inference (e.g., Baker et al. 1992; Jansen et al. 2003). For many applications in psychosocial research, ignorable missingness is often a plausible assumption. In such applications, we can use the model (15) to either test the MCAR assumption or obtain valid inference under MAR.

3 Application

We illustrate applications of the proposed approach with data from two real studies on sexual abuse and health research as well as from a simulated study. The simulated study allows us to

study the performance of the procedure in small to moderate sample sizes. We set the statistical significance at 0.05 for inference in all the examples.

3.1 Real Studies

Example 1—In many studies with vulnerable populations such as those in suicide, sexual abuse and substance use research (Achenbach, 2005; Bernstein et al. 2003; Heisel et al. 2006; Nelson et al. 1990; Shrier et al. 2005; Turner et al. 1998), a common strategy is to use informant sources to provide information about patients. In this example, we illustrate an application of the methodology to this line of research using data from a study that examines the relationship between patient and informant reports of childhood sexual abuse (Gamble et al. in press).

The study data come from a larger investigation examining the relationship between personality and suicidal behavior among depressed patients 50 years of age and older (Heisel et al. 2006). Among the 187 patients who completed a measure of childhood sexual abuse, 88 identified an informant source who completed a parallel version of the same measure. Sexual abuse was assessed by the Child Trauma Questionnaire (CTQ; Bernstein et al. 1998). Although CTQ is one of the most validated and reliable assessments of childhood sexual abuse (Bernstein and Fink, 1998; Bernstein et al. 2003; Scher et al. 2001), the concordance between the proband and informant reports on the CTQ had never been formally assessed, due in part to the lack of statistical methods. We compare the reliability in informant's report between the "Any" and "Severe" abuse categories, obtained by following the severity cut-point guidelines listed in the CTQ manual (Bernstein et al., 1998).

The analysis was based on the 88 subjects who identified an informant source to provide the informant data about their sexual abuse histories. Since every proband-informant pair of this subsample responded to the CTQ, there are no missing data. Shown in Table 1 are the sensitivity, specificity, PPV and NPV estimates computed for each of the two sexual abuse categories. As noted earlier in Section 2.2, it is also sensible to estimate and compare PPV and NPV, as these indicate whether the informant's information is reliable for assessing the proband's sexual abuse history, though all results should be interpreted with respect to the subgroup of subjects who were willing to provide informant's source. The results suggested differential reliability in informant's information between the two levels of sexual abuse severity. The values in Table 1 suggest that "Severe" abuse has a higher reliability than "Any" abuse. To formally assess statistical significance, we tested the null of no between-category difference for each of the accuracy indices. By applying the procedures in Section 2.2, we obtained the statistics (p-value) for testing such differences; $Q_n^2=21.13\ (<0.001)$ for sensitivity, $Q_n^2=20.57\ (<0.001)$ for PPV and $Q_n^2=9.36\ (0.002)$ for NPV. There was no significant difference for specificity.

Since there was no missing data in either the reference standard or the test outcome, methods based on the popular GLMM and GEE can be applied. For example, for inference about the sensitivities of informant's information for the two levels of abuse severity, ψ_k ($k = 1$ for "Any" and 2 for "Severe"), we modeled y_{ki} as a function of z_{ki} using either GLMM or GEE as follows:

$$\begin{aligned}
\text{GLMM} \quad & :y_{ki}\sim\text{i.d. Bernoulli}\,(\Pr(y_{ki}=1\,|\,z_{ki})), \quad 1\le i\le n, \quad 1\le k\le 2,\\
& :\text{logit}\,(\Pr(y_{ki}=1\,|\,z_{ki}))=\beta_0+\beta_1 z_{ki}+\beta_2 I_{\{k=2\}}+\beta_3 z_{ki} I_{\{k=2\}}+b_i, \quad b_i\sim N(0,\sigma_b^2),\\
\text{GEE} \quad & :\text{logit}\,(\Pr(y_{ki}=1\,|\,z_{ki}))=\beta_0+\beta_1 z_{ki}+\beta_2 I_{\{k=1\}}+\beta_3 z_{ki} I_{\{k=1\}}, \quad 1\le i\le n, \quad 1\le k\le 2,
\end{aligned} \tag{16}$$

where $I_{\{\cdot\}}$ denotes a set indicator, Bernoulli(p) indicates a Bernoulli distribution with the probability of success p, and i.d. stands for *independently distributed* (random variable). Under either model in (16), the sensitivity ψ_k was given by:

$$\psi_k=\frac{\exp\left(\Pr\left(y_{ki}=1\,\middle|\,z_{ki}=1\right)\right)}{1+\exp\left(\Pr\left(y_{ki}=1\,\middle|\,z_{ki}=1\right)\right)}=\frac{\exp\left(\beta_0+\beta_1+(\beta_2+\beta_3)\,I_{\{k=2\}}\right)}{1+\exp\left(\beta_0+\beta_1+(\beta_2+\beta_3)\,I_{\{k=2\}}\right)},\ 1\le k\le 2.$$

We tested the null, $H_0 : \beta_2+\beta_3 = 0$, to see if there was any differential sensitivity between "Severe" and "Any" abuse.

Similar models were constructed for the other indices. For example, by using the following GLMM and GEE,

$$\begin{aligned}
\text{GLMM}\quad &: y_{ki}\sim\text{i.d. Bernoulli}\left(\Pr\left(y_{ki}=0\,\middle|\,z_{ki}\right)\right),\quad 1\le i\le n,\quad 1\le k\le 2,\\
&:\text{logit}\left(\Pr\left(y_{ki}=0\,\middle|\,z_{ki}\right)\right)=\gamma_0+\gamma_1 z_{ki}+\gamma_2 I_{\{k=2\}}+\gamma_3 z_{ki} I_{\{k=2\}}+c_i,\quad c_i\sim N\left(0,\sigma_c^2\right),\\
\text{GEE}\quad &:\text{logit}\left(\Pr\left(y_{ki}=0\,\middle|\,z_{ki}\right)\right)=\gamma_0+\gamma_1 z_{ki}+\gamma_2 I_{\{k=1\}}+\gamma_3 z_{ki} I_{\{k=1\}},\quad 1\le i\le n,\quad 1\le k\le 2,
\end{aligned}$$

we estimated the specificity φ_k for each abuse level k (k = 1;2),

$$\varphi_k=\frac{\exp\left(\Pr\left(y_{ki}=0\,\middle|\,z_{ki}=0\right)\right)}{1+\exp\left(\Pr\left(y_{ki}=0\,\middle|\,z_{ki}=0\right)\right)}=\frac{\exp\left(\gamma_0+\gamma_2 I_{\{k=2\}}\right)}{1+\exp\left(\gamma_0+\gamma_2 I_{\{k=2\}}\right)},\ 1\le k\le 2,$$

and tested the null, $H_0 : \gamma_2 = 0$ to examine potential difference between φ_k.

Note that the GEE application to the current context requires the use of the working independence correlation structure to ensure consistent estimation of model parameters (e.g., Pepe and Anderson, 1994). In this case, it is readily checked that the estimating equations are readily solved to yield the same estimates as the proposed approach.

Shown in Table 1 are the estimates of the four indices based on the GLMM, obtained by the GLIMMIX procedure in SAS (SAS Institute, 2006). The GLMM estimates are quite close to their distribution-free counterparts except for the sensitivity estimate for the "Severe" abuse. It seems that the normal assumption for the random effort may not be appropriate for modeling the correlation between the informant's responses for the two abuse categories. By using linear contrasts, we obtained χ^2 statistics, $G_n^2\,(\text{p}-\text{values})$, for comparing the two abuse categories; $G_n^2=6.75\ (0.019)$ for sensitivity, $G_n^2=7.46\ (0.001)$ for PPV and $Q_n^2=8.69\ (0.004)$ for NPV. Again, the difference was not significant for specificity.

Example 2—Studies on sexual health and HIV prevention rely almost exclusively on retrospective self-report data to capture information on individuals' associated sexual behaviors, such as frequency of condom use. As such self-reports delve into very personal and private aspects of an individual's life, their accuracy is of growing concern when assessing treatment effects in prevention studies based on behavioral modifications (Catania et al. 1995; Kauth et al. 1991; Weinhardt et al. 1998). Various methods have been proposed and compared to improve the quality and reliability of such retrospective self-report data (Catania et al. 1995; Coxon, 1999; Des Jarlais et al. 1999; Locke et al. 1992; Metzger et al. 2000; Turner et al. 1998; de VincenziI, 1994; Graham et al. 2003; Jaccard and Wan, 1995; Kauth et al. 1991; Lagarde et al. 1995; Morrison-Beedy et al. 2006; Schroder et al. 2003; Weinhardt et al. 1998). In this example, we utilize the proposed approach to investigate whether reliability of

self-report on unprotected vaginal sex (UnPVS) is associated with the amount of such sex reported (Morrison-Beedy et al. 2006).

In this study, 160 adolescent girls monitored their behavior with a daily diary, and returned for assessment after three months. We used the daily diary as a reference standard for assessing accuracy of retrospective recall. Although a daily diary may not be 100% accurate, such a contemporaneous monitoring strategy addresses some of the key limitations of retrospective assessment, such as recall bias (Graham et al. 2003; Jaccard et al. 2002; Reading, 1983; Shrier et al. 2005).

To help assess reliability of self-report on UnPVS, we created five categories for evidence of UnPVS based on the % of UnPVS reported over the three month period; 5%, 10%, 30%, 40% and 50%. For a given category defined by x%, a subject was defined by the reference standard or test as having UnPVS over the period if she reported at least x% of UnPVS in the diary or the retrospective report. To apply the proposed approach, let z_{ki} (y_{ki}) denote the status of UnPVS as indicated by the daily diary (retrospective report) for the ith subject in the study based on the kth category ($1 \leq k \leq 5$).

There was about 10% missing data from both the diary and retrospective assessment data. We modeled the missingness according to (15) by including behavioral intention, condom use attitude, depression, HIV knowledge, race, and incidents of protected as well as unprotected vaginal sex at baseline (Morrison-Beedy et al. 2006) in the covariate vector $\mathbf{x}_i$. A backward elimination procedure helped trim the model to only one covariate, HIV knowledge, with a marginally significant p-value = 0.08.

Shown in Table 2 are the IPW-based estimates of the four measures of accuracy for the five UnPVS categories. As in Example 1, PPV and NPV indicate the degree of accuracy when retrospective reports on UnPVS are used to assess the actual practice of this unsafe sex act for this study population as benchmarked by the daily diary. The higher sensitivity and PPV estimates suggest that self-reports are generally quite reliable for detecting UnPVS in this study population. All estimates seem to initially rise and then fall with a peak at 30% as the % of UnPVS reported by the daily diary varied from 5% to 50%.

Shown in Table 3 are the p-values from testing the null of no difference across the five categories based on the inference procedure in Section 2.3. The results indicate significant differences for Specificity and NPV. For these two indices, Table 3 provides the p-values for comparing the 30% category with each of the other categories. The results indicate that specificity and NPV rise significantly as the % of UnPVS increases and then level off after 30%. Although estimates of sensitivity and PPV exhibit a similar pattern, the differences are not significant.

3.2 A Simulation Study

We conducted a limited simulation study to examine the empirical type I error rate for testing the null of equal sensitivities across different diagnostic criteria as well as empirical power for the alternative hypothesis of varying sensitivities, with three reference categories and four sample sizes–50, 100, 150 and 2000–under complete as well as missing data modeled by the MCAR and MAR mechanisms. We simulated (z_{ki}; y_{ki}) according to the following distributions:

$$z_{ki} \sim Bi(p_k), \quad y_{ki} \mid z_{ki}=1 \sim Bi(\phi_k), \quad y_{ki} \mid z_{ki}=0 \sim Bi(1-\psi_k), \quad 1 \leq k \leq 3,$$

where B_i (p) is a Bernoulli distribution with the probability of success p and

$$p_k=\Pr(z_{ki}=1),\quad \phi_k=\Pr(y_{ki}=1\mid z_{ki}=1),\quad \psi_k=\Pr(y_{ki}=0\mid z_{ki}=0),\quad 1\le k\le 3.$$

We set the prevalence p_k = 0:65 and test specificity ψ_k = 0:7 ($1 \le k \le 3$). To create correlated responses, we set ρ_k = Pr ($z_{ki} = z_{(k+1)i}$ = 1) = 0:55 (correlation between consecutive z_{ki}'s for k = 1; 2). For investigating type I error rates, we generated (z_{ki}; y_{ki}) under the null of equal sensitivity across the three diagnostic criteria, $H_0 : \phi_1 = \phi_2 = \phi_3$ = 0:8, while for examining power, we generated (z_{ki}; y_{ki}) from the alternative, $H_a : \phi_1$ = 0:8; $\phi_2 = \phi_3$ = 0:9.

We simulated the missing response for the MAR model with about 25% missing data according to the generalized logit model in (15), with p_{ij} given by:

$$\begin{aligned} p_{ij} &= \exp(\eta_{0j}+\eta_{xj}x_i),\quad 1\le i\le n,\quad 1\le j\le 4,\\ \eta_{01} &= \eta_{x1}=0,\quad \eta=(\eta_{02},\eta_{03},\eta_{04},\eta_{x2},\eta_{x3},\eta_{x4})^{\top}. \end{aligned}$$

We set η_{xj} = 3 ($2 \le j \le 4$) for the dependence of missingness on the covariate x_i, which was simulated from a standard normal variate. To yield about 25% missing response, we determined η_{0j} by solving the following equation:

$$\sum_{i=1}^{n} p_{i1}(x_i,\eta)=\sum_{i=1}^{n}\frac{1}{1+\sum_{j=2}^{4}\exp(\eta_{0j}+\eta_{xj}x_i)}=0.75n. \tag{17}$$

The same process was used for simulating missing response under MCAR by setting η_{xj} = 0 and solving for η_{0j} in (17). Sensitivity estimates of ϕ_k and their asymptotic variance estimates were obtained based on the results in (13). The empirical type I error rate (power) for testing the null H_0 (alternative H_a) was calculated according to

$\widehat{\alpha}=\frac{1}{M}\sum_{j=1}^{M} I_{\{Q_{nj}^2\ge q_{0.95}\}}$ $(\widehat{\omega}=\frac{1}{M}\sum_{j=1}^{M} I_{\{Q_{nj}^2\ge q_{0.95}\}})$, where M indicates the Monte Carlo (MC) sample size, Q_{nj}^2 denotes the test statistic Q_n^2 in Section 2.1 from the jth MC replication constructed based on the data simulated under H_0 (H_a), and $q_{0:95}$ designates the 95th percentile of the χ^2 distribution with 2 degrees of freedom.

Shown in Table 4 are the averaged estimates of sensitivity along with the averaged asymptotic standard errors and empirical type I error rates under H_0, based on 1,000 Monte Carlo (MC) replications. Sensitivity estimates are quite close to the true parameter values, even for sample size 50, though the empirical type I error rates are a bit upwardly biased for sample sizes ≤ 100. The proposed estimates for the two missing data cases seem to perform well across all the sample sizes relative to the complete data case.

Power estimates based on 1,000 MC replications for the alternative H_a are shown in Table 5, together with the averaged estimates of sensitivity and asymptotic standard errors. As in Table 4, sensitivity estimates are quite good. Power estimates increased as a function of sample size, reaching the value 1 for n = 2000. Power is not great for sample sizes ≤ 150 as binary response is notorious for low power as compared to its continuous counterpart.

As expected, maximum power occurred for the complete data case, with the two missing data models yielding reduced power. Between the two missing data cases, more power was achieved under MAR. This may not be surprising since unlike MCAR, MAR attempts to augment each ith observed response with the weight functions π_{zyi} and π_{zi} to "statistically recover" the missing

data in estimating the sensitivities. The additional information provided by the observed responses in modeling π_{zyi} and π_{zi} helps to increase power for the MAR model.

4 Future Research

We developed an approach for inference when comparing multiple diagnostic tests with varying diagnostic criteria. Although relatively infrequent in biomedical applications, the need for such comparisons arises quite often in the behavioral and social sciences. As methods for inference about such comparisons in the presence of missing data are currently lacking, the proposed methodology fills this important gap in the literature.

The methodologic issues considered here are quite different from ROC curves analysis. Although different sensitivity and specificity estimates are examined in ROC analysis, these estimates arise from varying the cut-point in dichotomizing an underlying continuous test outcome for detecting a common diagnostic condition, rather than by different diagnostic criteria that affect both the test outcome and the reference standard as in the current setting.

We focused on non-parametric inference to address missing data, and in particular articulated a form of the MAR assumption within the current context and discussed inference using a class of IPW estimates. Our approach reduces to the WGEE estimate under a single diagnostic criterion with no missing data in the reference standard (e.g., Robins et al. 1995).

Alternatively, likelihood based or Bayesian inference may also be considered (e.g. Johnson and Gastwirth, 1991; Mendoza-Blanco et al. 1996; Prentice, 1988; Leisenring et al. 2000; Chaganty and Joe, 2004); the latter is especially appropriate when one wishes to incorporate information from other prior testing data. However, such approaches are much more complicated when modeling multiple correlated binary outcomes even under complete data. We have opted for the nonparametric approach as it affords simpler and more intuitive estimates.

Important weaknesses of the approach include its inability to control for covariates and to handle missing data when the data occur differentially across the different diagnostic criteria rather than following the same pattern (as assumed in the current development). Work is currently underway to address these limitations.

Appendix

Appendix

We show that it is not possible to model the missingness of $\mathbf{z}_i$ as dependent on $\mathbf{y}_i$ and vice versa in addition to $\mathbf{x}_i$ under MAR. For notational brevity, we also suppress the dependence on $\mathbf{x}_i$.

Suppose that on the contrary such a model existed. Then, we would have:

$$\begin{aligned}\Pr[r_{zi}=1 \mid r_{yi}=1, \mathbf{z}_i, \mathbf{y}_i] &= \Pr[r_{zi}=1 \mid r_{yi}=1, \mathbf{y}_i],\\ \Pr[r_{yi}=1 \mid r_{zi}=1, \mathbf{z}_i, \mathbf{y}_i] &= \Pr[r_{yi}=1 \mid r_{\mathbf{z}i}=1, \mathbf{z}_i].\end{aligned} \tag{18}$$

Under MAR, the probabilities of missing response depend only on observed data. It follows that $\Pr[r_{zi} = 1; r_{yi} = 0 \mid \mathbf{z}_i; \mathbf{y}_i]$ is a function of $\mathbf{z}_i$ and $\Pr[r_{zi} = 0; r_{yi} = 1 \mid \mathbf{z}_i; \mathbf{y}_i]$ a function of $\mathbf{y}_i$ only. Denote them as $f(\mathbf{z}_i)$ and $g(\mathbf{y}_i)$, respectively. Then, $\Pr[r_{zi} = 1; r_{yi} = 1 \mid \mathbf{z}_i; \mathbf{y}_i] = 1 - f(\mathbf{z}_i) - g(\mathbf{y}_i)$. It follows that

$$\begin{aligned}\Pr[r_{zi}=1 \mid \mathbf{z}_i, \mathbf{y}_i] &= 1 - f(\mathbf{z}_i) - g(\mathbf{y}_i) + f(\mathbf{z}_i) = 1 - g(\mathbf{y}_i),\\ \Pr[r_{yi}=1 \mid r_{zi}=1, \mathbf{z}_i, \mathbf{y}_i] &= \frac{1-f(\mathbf{z}_i)-g(\mathbf{y}_i)}{1-g(\mathbf{y}_i)}.\end{aligned}$$

It follows from (18) that $\frac{1 - f(\mathbf{z}_i) - g(\mathbf{y}_i)}{1 - g(\mathbf{y}_i)}$ is a function of $\mathbf{z}_i$ only. Thus, $g(\mathbf{y}_i)$ must be a constant. Likewise, $f(\mathbf{z}_i)$ must be a constant. These contradict the MAR assumption.

Acknowledgment

This research was supported in part by NIH grants R01-MH60285, R01-DA012249, 1 UL1 RR024160, and R24-MH071604. We are also deeply indebted to Ms. Bliss-Clark in the Department of Biostatistics and Computational Biology at the University of Rochester, two anonymous reviewers, an associate editor and Editor Azen for helpful and constructive comments that led to great improvement in the presentation.

References

1. Achenbach TM, Krukowski RA, Dumenci L, Ivanova MY. Assessment of adult psychopathology: meta-analyses and implications of cross-informant correlations. Psychological Bulletin 2005;131:361–382. [PubMed: 15869333]
2. Ahn C. Statistical methods for the estimation of sensitivity and specificity of site-specific diagnostic tests. Journal of Periodontal Research 1997;32:351–354. [PubMed: 9210088]
3. Baker SG, Rosenberger WF, DerSimonian R. Closed-form estimates for missing counts in two-way contingency tables. Statistics in Medicine 1992;11:643–657. [PubMed: 1594807]
4. Bernstein, DP.; Fink, L. CTQ Childhood Trauma Questionnaire. A retrospective self-report. San Antonio, TX: Harcourt Brace and Company; 1998.
5. Bernstein D, Fink L, Handelsman L, Foote J, Lovejoy M, Wenzel K, Sapareto E, Ruggiero J. Initial reliability and validity of a new retrospective measure of child abuse and neglect. American Journal of Psychiatry 1994;151:1132–1136. [PubMed: 8037246]
6. Bernstein DP, Stein JA, Newcomb MD, Walker E, Pogge D, Ahluvalia T, Stokes J, Handelsman L, Medrano M, Desmond D, Zule W. Development and validation of a brief screening version of the Childhood Trauma Questionnaire. Child Abuse and Neglect 2003;27:169–190. [PubMed: 12615092]
7. Catania JA, Binson D, Dolcini MM, et al. Risk factors for HIV and other sexually transmitted diseases and prevention practices among US heterosexual adults: Changes from 1990 to 1992. American Journal of Public Health 1995;85:1492–1499. [PubMed: 7485660]
8. Chaganty R, Joe H. Efficiency of generalised estimating equations for binary responses. J. R. Statist. Soc. B 66:851–860.
9. Coxon AP. Parallel accounts? Discrepancies between self-report (diary) and recall (questionnaire) measures of the same sexual behavior. AIDS Care 1999;11:221–234. [PubMed: 10474624]
10. Cross GD, Schalla WO, Hancock JS, et al. Analytic sensitivity and specificity of Enzyme Immunoassay results in testing for Human Immunodeficiency Virus type 1 antibody. Arch Pathol Lab Med 1992;116:477–481. [PubMed: 1580749]
11. Des Jarlais DC, Paone D, Milliken J, et al. Audio-computer interviewing to measure risk behavior for HIV among injecting drug users: A quasi-randomized trial. The Lancet 1999;353:1657–1661.
12. de VincenziI. A longitudinal study of human immunodeficiency virus transmission by heterosexual partners. European Study Group on Heterosexual Transmission of HIV. N. Engl. J. Med 1994;331:341–346. [PubMed: 8028613]
13. Fleiss, JL.; Levin, B.; Paik, MC. Statistical methods for rates and proportions. Vol. 3rd ed.. New York: Wiley;
14. Gamble SA, Talbot NL, Conner KR, Tu XM, Franus N, Beckman AM, Ma Y, Duberstein PR. Concordance about childhood sexual abuse among depressed patients and their family and friends. Archives of Suicide Research. in press

15. Genc Y, Gokmen D, Tuccar E, Yagmurlu B. Estimation of sensitivity and specificity for clustered data. Turk J Med Sci 2005;35:21–24.
16. Graham CA, Catania JA, Brand R, Canchola JA. Recalling sexual behavior: A methodological analysis of memory recall bias via interview using the diary as a gold standard. Journal of Sex Research 2003;40:325–332. [PubMed: 14735406]
17. Heisel MJ, Duberstein PR, Conner KR, Franus N, Beckman A, Conwell Y. Personality and reports of suicide ideation among depressed adults 50 years of age or older. 2006Manuscript submitted for publication
18. Jaccard J, Wan CK. A paradigm for studying the accuracy of self-reports of risk behavior relevant to aids: Empirical perspectives on stability, recall bias, and transitory influences. Journal of Applied Social Psychology 1995;25:1831–1858.
19. Jaccard J, McDonald R, Wan CK, Dittus PJ, Quinlan S. The accuracy of self-reports of condom use and sexual behavior. Journal of Applied Psychology 2002;32:1863–1905.
20. Jansen I, Molenberghs G, Aerts M, Thijs H, Van Steen K. A local influence approach applied to binary data from a psychiatric study. Biometrics 2003;59:410–419. [PubMed: 12926726]
21. Johnson WO, Gastwirth JL. Bayesian inference for medical screening tests: Approximations useful for the Analysis of Acquired Immune Deficiency Syndrome'. J. R. Statist. Soc. B 1991;53:427–439.
22. Kauth MR, St. Lawrence JS, Kelly JA. Reliability of retrospective assessments of sexual HIV risk behavior: A comparison of biweekly, three-month, and twelve-month self-reports. AIDS Education & Prevention 1991;3:207–214. [PubMed: 1834142]
23. Kowalski J, Tu XM, Jia MA, Pagano M. A comparative meta-analysis on the variability in test performance among FDA licensed enzyme immunosorbent assays for HIV antibody testing. Journal of Clinical Epidemiology 2001;54:448–461. [PubMed: 11337207]
24. Kowalski, J.; Tu, XM. Modern Applied U Statistics. New York: Wiley; 2007.
25. Lagarde E, Enel C, Pison G. Reliability of reports of sexual behavior: a study of married couples in rural West Africa. Am J Epidemiol 1995;141:1194–1200. [PubMed: 7771458]
26. Lee E, Dubin N. Estimation and sample size considerations for clustered binary responses. Stat in Med 1994;13:1241–1252. [PubMed: 7973205]
27. Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics 2000;56:345–351. [PubMed: 10877288]
28. Lipschitz DS, Bernstein DP, Winegar RK, Southwick SM. Hospitalized adolescents' reports of sexual and physical abuse: a comparison of two self-report measures. Journal of Traumatic Stress 1999;12:641–654. [PubMed: 10646182]
29. Locke SE, Kowaloff HB, Hoff RG, et al. Computer-based interview for screening blood donors for risk of HIV transmission. Journal of the American Medical Association 1992;268:1301–1305. [PubMed: 1507376]
30. Mendoza-Blanco JR, Tu XM, Iyengar S. Bayesian inference on prevalence using a missing-data approach with simulation-based techniques: application to HIV screening. Statistics in Medicine 1996;15:2161–2176. [PubMed: 8910961]
31. Metzger DS, Koblin B, Turner C, et al. Randomized controlled trial of audio computer-assisted self-interviewing: Utility and acceptability in longitudinal studies. American Journal of Epidemiology 2000;152:99–106. [PubMed: 10909945]
32. Morrison-Beedy D, Carey MP, Tu XM. Accuracy of audio computer-assisted self-interviewing (ACASI) and self-administered questionnaires for the assessment of sexual behavior. AIDS Behav 2006;10:541–552. [PubMed: 16721506]
33. Nelson LM, Longstreth WT, Koepsell TD, VanBelle G. Proxy respondents in epidemiologic research. Epidemiologic Reviews 1990;12:71–86. [PubMed: 2286227]
34. Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics 1988;44:321–327.
35. Reading AE. A comparison of the accuracy and reactivity of methods of monitoring male sexual behavior. Journal of Behavioral Assessment 1983;5:11–23.
36. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 1995;90:106–121.r3

37. Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for re-peated outcomes with nonignorable nonresponse. Journal of the American Statistical Association 1998;93:1321–1339.
38. Rubin DB. Inference and missing data. Biometrika 1976;63:581–592.
39. SAS Institute. GLIMMIX procedure: an add-on in SAS 9.1 to SAS/STAT for the (32-bit) Windows platform. Cary, NC: SAS Institute Inc.; 2006.
40. Scher CD, Stein MB, Asmundson GJG, McCreary DR, Forde DR. The Childhood Trauma Questionnaire in a community sample: psychometric properties and normative data. Journal of Traumatic Stress 2001;14:843–857. [PubMed: 11776429]
41. Schroder KEE, Carey MP, Vanable PA. Methodological challenges in research on sexual risk behavior: II. Accuracy of self-reports. Annals of Behavioral Medicine 2003;26:104–123. [PubMed: 14534028]
42. Shrier LA, Shih M, Beardslee WR. Comparison of momentary sampling with diary and retrospective self-report methods of measurement. Pediatrics 2005;115:573–581.
43. Talbot JA, Talbot NL, Tu XM. Shame-proneness as a diathesis for dissociation in women with histories of childhood sexual abuse. Journal of Traumatic Stress 2004;17:445–448. [PubMed: 15633925]
44. Taylor RN, Hearn TL, Schalla WO, et al. Indirect immunofluorescence test performance and questionnaire results from the Centers for Disease Control Model Performance Evaluation Program for Human Immunodeficiency Virus Type 1 testing. J Clin Microbiol 1990;28:1799–1807. [PubMed: 2168439]
45. Tu XM, Litvak E, Pagano M. Issues in HIV screening programs. American Journal of Epidemiology 1992;136:244–255. [PubMed: 1415146]
46. Tu XM, Litvak E, Pagano M. Screening tests: Can we get more by doing less. Statistics in Medicine 1994;13:1905–1919. [PubMed: 7846399]
47. Tu XM, Litvak E, Pagano M. On the informativeness and accuracy of pooled testing in estimating prevalence for a rare disease: Application to HIV screening. Biometrika 1995;82:287–297.
48. Turner CF, Ku L, Rogers SM, et al. Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. Science 1998 May;:867–873. [PubMed: 9572724]
49. Weinhardt LS, Forsyth AD, Carey MP, Jaworski BC, Durant LE. Reliability and validity of self-report measures of HIV-related sexual behavior: Progress since 1990 and recommendations for research and practice. Archives of Sexual Behavior 1998;27:155–180. [PubMed: 9562899]
50. Yasemin Geng DG, Tuccar Ersoz, Yagmurlu Banu. Estimation of sensitivity and specificity for clustered data. Turk J Med Sci 2005;35:21–24.

Table 1

Sensitivity, specificity, PPV and NPV estimates for two categories of sexual abuse history reported by probands and informants in Example 1 based on the proposed (GEE) and GLMM approaches.

Comparisons of Estimates by Categories of Sexual Abuse History Proposed (GEE) / GLMM				
Abuse category	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Any abuse	0.51 / 0.50	0.91 / 0.92	0.84 / 0.84	0.68 / 0.68
Severe abuse	0.64 / 0.55	0.94 / 0.94	0.78 / 0.77	0.89 / 0.90

Table 2

IPW-based estimates of sensitivity, specificity, PPV and NPV as a function of % of unprotected vaginal sex reported by daily diary over a three month period as assessed by two methods of retrospective reporting in Example 2.

% of unprotected vaginal sex reported	Sensitivity	Specificity	PPV	NPV
5%	0.87	0.79	0.95	0.59
10%	0.88	0.76	0.93	0.64
30%	0.89	0.96	0.98	0.80
40%	0.82	0.84	0.89	0.75
50%	0.77	0.92	0.92	0.77

Table 3

p-values for testing the null of no difference across the five diagnostic categories and the null hypotheses for pairwise comparisons of 30% vs. each of the other categories for the accuracy indices in Example 2.

Comparison	Sensitivity	Specificity	PPV	NPV
No diff. across five categories	0.79	<0.001	0.44	<0.01
5% vs. 30%	0.73	<0.01	0.61	<0.01
10% vs. 30%	0.86	<0.001	0.37	<0.01
40% vs. 30%	0.33	0.09	0.21	0.48
50% vs. 30%	0.24	0.70	0.56	0.77

Table 4

Averaged estimates of sensivity over 1,000 MC replications along with asymptotic standard errors and empirical type I error rates under the null hypothesis.

Parameter	Sample size											
	***n* = 50**			***n* = 100**			***n* = 150**			***n* = 2000**		
	Complete data											
$\hat{\phi}$	.800	.803	.799	.800	.801	.799	.800	.802	.799	.800	.800	.800
$\hat{\Sigma}_{\phi}$	.005	.005	.005	.002	.002	.002	.002	.002	.002	.0001	.0001	.0001
$\hat{\alpha}$	.069			.061			.053			.045		
	Missing data under MCAR											
$\hat{\phi}$	.799	.798	.799	.801	.799	.802	.799	.800	.799	.800	.800	.800
$\hat{\Sigma}_{\phi}$	.009	.008	.007	.004	.004	.004	.003	.003	.003	.0002	.0002	.0002
$\hat{\alpha}$	.075			.06			.056			.046		
	Missing data under MAR											
$\hat{\phi}$	.800	.804	.800	.801	.801	.801	.801	.801	.800	.800	.800	.800
$\hat{\Sigma}_{\phi}$	.007	.006	.007	.003	.003	.003	.002	.002	.002	.0001	.0001	.0001
$\hat{\alpha}$	.069			.057			.051			.046		

Table 5

Averaged estimates of sensivity over 1,000 MC replications along with asymptotic standard errors and empirical power under the alternative hypothesis.

Parameter	**Sample size**											
	***n* = 50**			***n* = 100**			***n* = 150**			***n* = 2000**		
	Complete data											
$\hat{\phi}$	.799	.903	.899	.800	.902	.901	.801	.900	.900	.799	.900	.900
$\hat{\Sigma}_{\phi}$	.005	.003	.003	.002	.001	.001	.002	.001	.001	.0002	.0001	.0001
$\hat{\omega}$	.246			.372			.502			1		
	Missing data under MCAR											
$\hat{\phi}$	.799	.901	.898	.801	.901	.900	.800	.898	.900	.799	.900	.900
$\hat{\Sigma}_{\phi}$	.01	.005	.006	.005	.003	.003	.003	.002	.001	.0003	.0002	.0002
$\hat{\omega}$	.193			.244			.286			1		
	Missing data under MAR											
$\hat{\phi}$	.799	.902	.900	.802	.903	.901	.801	.901	.902	.799	.900	.900
$\hat{\Sigma}_{\phi}$	.007	.004	.004	.003	.002	.002	.002	.001	.001	.0002	.0001	.0001
$\hat{\omega}$	.203			.278			.383			1		