HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2016 April 30.

Published in final edited form as:

Stat Med. 2015 April 30; 34(9): 1605–1620. doi:10.1002/sim.6440.

Binary Regression with Differentially Misclassified Response and Exposure Variables

Li Tang[*], **Robert H. Lyles**, **Caroline C. King**, **David D. Celentano**, and **Yungtai Lo**

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, The United States

Li Tang: li.tang@stjude.org

Abstract

Misclassification is a long-standing statistical problem in epidemiology. In many real studies, either an exposure or a response variable or both may be misclassified. As such, potential threats to the validity of the analytic results (e.g., estimates of odds ratios) that stem from misclassification are widely discussed in the literature. Much of the discussion has been restricted to the nondifferential case, in which misclassification rates for a particular variable are assumed not to depend on other variables. However, complex differential misclassification patterns are common in practice, as we illustrate here using bacterial vaginosis (BV) and Trichomoniasis data from the HIV Epidemiology Research Study (HERS). Therefore, clear illustrations of valid and accessible methods that deal with complex misclassification are still in high demand. We formulate a maximum likelihood (ML) framework that allows flexible modeling of misclassification in both the response and a key binary exposure variable, while adjusting for other covariates via logistic regression. The approach emphasizes the use of internal validation data in order to evaluate the underlying misclassification mechanisms. Data-driven simulations show that the proposed ML analysis outperforms less flexible approaches that fail to appropriately account for complex misclassification patterns. The value and validity of the method is further demonstrated through a comprehensive analysis of the HERS example data.

Keywords

Likelihood; logistic regressions; misclassification; odds ratio

1. Introduction

For many epidemiologic studies, reliably estimating effects of exposures on a health outcome is of primary interest. It has been noted in a wide range of study contexts [1–8] that the problem of measurement error may arise due to fallibility in measurement tools,

[*]Correspondence to: Li Tang, Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, The United States.

Author affiliations: Dept. of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee (L. Tang); Dept. of Biostatistics and Bioinformatics, The Rollins School of Public Health, Emory University (R.H. Lyles); Division of Reproductive Health, Centers for Disease Control and Prevention, Atlanta, Georgia (C.C. King); Dept. of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York (Y. Lo); Dept. of Epidemiology, The Johns Hopkins Bloomberg School of Public Health (D.D. Celentano).

hindering investigators in the effort to report valid estimates of effect. In addition to these references, many general discussions about the impact of measurement error on estimating health effects and approaches to combat it can be found in the literature [e.g., 9–14].

When referring to measurement error in binary or categorical variables, the term misclassification is commonly used, and it has been sometimes considered as a distinct concept. For example, to assess disease status, epidemiologic studies sometimes use convenient but error-prone diagnostic tests, which may lead to false positive and negative diagnoses. There is rich literature on methods dealing with misclassification in exposure or response variables, often focusing on nondifferential misclassification or invoking sensitivity analyses when errors may be differential [e.g., 15–18]. However, the value of internal validation data for elucidating and adjusting for exposure misclassification that is differential with respect to disease status has been made very clear [e.g., 11, 19–22]. While some publications advocate the use of internal validation data to evaluate a more complex misclassification process, relatively few emphasize practical guidance on how to implement the analysis. Our current work seeks to address aspects of the gap between theoretical considerations and real life applications when one encounters potentially complex patterns of binary variable misclassification.

Here we consider a motivating example arising in the HIV Epidemiology Research Study (HERS) [23], where both a binary exposure variable X (Trichomoniasis (Trich) status) and a binary response variable Y (bacterial vaginosis (BV) status) are subject to misclassification. The presence of outcome and/or exposure misclassification complicates regression analyses and can lead to distorted effect estimates (e.g., estimated odds ratios (ORs)). Moreover, the misclassification process with respect to BV in the HERS data has been found to be quite complex, making one doubt both the common assumption of nondifferentiality and typical differential mechanisms in which false positive and negative rates depend only on a single variable. Exploration into the misclassification of BV status further revealed that inappropriately assuming nondifferentiality can lead to corrected OR estimates as or more biased than those arising from “naïve” analyses that ignore the problem altogether [24].

In the following section, we first introduce the HERS data and discuss analytical questions arising with regard to logistic regression analysis of cross-sectional data involving Trich and BV status. When misclassification is present in both X and Y, prior discussions in the literature mainly target 2-by-2 tables or generalized tables [25, 26]. However, in practice it is very common to seek direct adjustment for covariates via regression modeling, as in our motivating example. Thus, we propose a unified likelihood framework to incorporate the main/internal validation design and assess complex misclassification mechanisms involving both binary variables. We present data-driven simulations mimicking the HERS example to demonstrate the benefits of the main/internal validation study design and corresponding likelihood-based methods, relative to a naïve analysis and an alternative approach that adjusts for misclassification based on overly simplistic assumed mechanisms. We also conduct extensive simulations under various settings to evaluate the performance of the proposed approach.

2. Materials and methods

2.1 Motivation

We consider a cross-sectional analysis of data on bacterial vaginosis (BV) as the response Y and Trichomoniasis (TRICH) status as the exposure X for women in the HIV Epidemiology Research Study (HERS) at the 4th semi-annual visit (between 09/1994 and 10/1996), given misclassification of BV status previously documented at that visit [24]. The primary goal of the analysis is to model the association of BV with TRICH, controlling for other subject-specific characteristics. Existing literature suggests a positive relationship between these two conditions [27], highlighting the motivation for assessing their prevalences and association. Unfortunately, in practice BV and TRICH may be diagnosed via error-prone tests. In HERS, these consist of a clinical method (hereafter referred to as "CLIN") for BV assessment based on a modified version of Amsel's criteria [28], and a wet mount testing procedure [29] for detecting TRICH. Although such clinically-based methods are convenient and cost-saving, they are subject to yielding false positives or negatives. In large-scale observational studies like HERS that seek insight about disease associations, the presence of misclassification in such diagnoses jeopardizes the validity of adjusted odds ratio estimates and corresponding statistical inferences.

A total of 904 women with complete data on BV, TRICH and other risk factors at the 4th visit were considered. Among them, 61.7% were black, 67.4% were HIV positive and 52% were intravenous drug users. The median age at enrollment was 37 years. Using culture testing (the arguable gold standard), 18% of women were diagnosed with TRICH. In contrast, only 7.6% were TRICH-positive based on wet mounts. Similarly, 40.3% of women were BV positive via a gold standard laboratory-based ("LAB") method based on Nugent's criteria [30], compared to 24.5% based on the CLIN method. Estimated crude sensitivities for the wet mount and CLIN methods were 37.8% and 51.7% respectively, while the specificities were 93.9% and 99.0%, indicating that both error-prone methods were highly specific but not sensitive.

We refer to the binary random variables corresponding to the gold standard TRICH and BV assessments as X and Y, respectively. The error-prone assessments are conversely labeled as X^* and Y^*. In HERS, all four assessments (X, X^*, Y, Y^*) were obtained for each participant at Visit 4. The complete data on X and Y provide an ideal data example to illustrate the performance of the proposed approach, since we can compare the results of a standard logistic regression applied to the gold standard data with those of a naïve analysis based on (X^*, Y^*) and with ML analyses incorporating only a portion of the (X, Y) data into an internal validation sample.

We first consider a multiple logistic model of interest on all subjects using Y and X as the response and predictor variables, referred to as the "Ideal Analysis". Preliminary model selection suggested that TRICH status, age, race (black vs. others), HIV risk cohort (RISKCHRT: intravenous drug use (IDU) vs. sexual) and HIV status (HIVPOS: positive vs. negative) are important risk factors for BV, as represented in model (1):

$$logit[\Pr(Y=1)]=\beta_0+\beta_1 X+\beta_2 AGE+\beta_3 RACE+\beta_4 RISKCHRT+\beta_5 HIVPOS. \quad (1)$$

We then fit the same model replacing X and Y with X^* and Y^*, denoted as the "Naïve Analysis". As seen in Table 1, the two analyses differ markedly in magnitudes of the estimated OR for TRICH (2.41 with 95% CI (1.66, 3.50) for ideal vs. 3.44 with 95% CI (2.03, 5.84) for naïve) and HIV risk cohort (1.37 with 95% CI (1.03, 1.83) for ideal vs. 2.45 with 95% CI (1.74, 3.45) for naïve). The estimated ORs for HIV status differ in directionality (1.25 with 95% CI (0.93, 1.69) for ideal vs. 0.73 with 95% CI (0.52, 1.03) for naïve). These discrepancies clearly indicate that without appropriately taking misclassification into account, estimates of effect and corresponding inferences that rely only upon X^* and/or Y^* can be invalid.

2.2 Misclassification patterns and joint likelihood formulation

Consider a cross-sectional study with n subjects. In the absence of misclassification, assume that we seek to fit a standard logistic regression model as follows:

$$logit[\Pr(Y=1|X, C_1, C_2 \ldots C_P)]=\beta_0+\sum_{p=1}^{P}\beta_p C_p+\beta_{P+1}X, \quad (2)$$

where Y is a potentially misclassified binary response variable, C_p (p=1,…P) denotes the pth covariate assumed to be measured accurately, and X stands for a binary predictor of interest that (like Y) is subject to misclassification. Henceforth, we divide the whole study sample into two parts. In the main study sample, instead of X and Y, mismeasured dichotomous exposure status X^* and disease status Y^* are observed. For the purpose of evaluating misclassification patterns, we assume that a completely random internal validation sample of size n_v is selected and gold standard measures of the response and exposure (Y and X) are made on those subjects. It follows that the main study sample size is $n_m=n-n_v$. We note in passing that some authors refer to the whole study sample as the "main study", which differs slightly from our use of the term. If we replace X and Y in eqn. (2) with the error-prone measures X^* and Y^*, estimates of the β's will be potentially biased (even in large samples). The magnitudes of the biases impacting these estimates depend on the diagnostic properties of the methods used to classify both X^* and Y^*.

2.2.1 Independent nondifferential misclassification—The assumption of nondifferential misclassification implies the belief that the common diagnostic parameters known as sensitivity (SE) and specificity (SP) are constants that do not vary based upon any subject-specific variables [e.g., 7]. For example, regarding diagnostic properties relating Y^* to Y in the nondifferential case, we define

$$SE_Y=\Pr(Y^*=1|Y=1) \text{ and } SP_Y=\Pr(Y^*=0|Y=0). \quad (3)$$

Similarly, when characterizing the method classifying X^*, nondifferentiality implies that

$$SE_X = \Pr(X^*=1|X=1) \text{ and } SP_X = \Pr(X^*=0|X=0). \quad (4)$$

The subscript notation used in (3) and (4) reflects the fact that SE and SP are viewed as constants independent of other information, such as disease status and prognostic factors. In other words, nondifferential misclassification for both variables corresponds to the following assumptions: $Pr(Y^* = 1|Y = 1) = Pr(Y^* = 1|Y = 1, X, \boldsymbol{C})$, $Pr(X^* = 1|X = 1) = Pr(X^* = 1|X = 1, Y, \boldsymbol{C})$. It follows that $Pr(Y^*, X^*, Y, X|\boldsymbol{C}) = Pr(Y^*|Y, X^*, X, \boldsymbol{C})\, Pr(X^*|X, Y, \boldsymbol{C})\, Pr(Y|X, \boldsymbol{C})\, Pr(X|\boldsymbol{C}) = Pr(Y^*|Y)\, Pr(X^*|X)\, Pr(Y|X, \boldsymbol{C})\, Pr(X|\boldsymbol{C})$, indicating the misclassification processes in X and Y are independent. This type of misclassification pattern is hereafter termed "independent nondifferential misclassification".

We set out to find expressions for the observed data likelihood under various misclassification assumptions. Let us first consider further the situation when both X and Y are assumed subject to independent nondifferential misclassification. By the rule of total probability, each independent observation in the main study contributes the following likelihood term:

$$\begin{aligned} Pr(Y^*=y^*, X^*=x^*|\boldsymbol{C}=\boldsymbol{c}) &= \sum_{y=0}^{y=1}\sum_{x=0}^{x=1} Pr(y^*, x^*, y, x|\boldsymbol{c}) \\ &= \sum_{y=0}^{y=1}\sum_{x=0}^{x=1} Pr(y^*|y)Pr(x^*|x)Pr(y|x, \boldsymbol{c})Pr(x|\boldsymbol{c}). \end{aligned} \quad (5)$$

The first and second terms in the last line of eqn. (5) represent the SE and SP of Y and X, while the third term reflects the primary model of interest defined in eqn. (2). The last term characterizes the association of X with other covariates $\boldsymbol{C}$. To facilitate the likelihood specification, we designate a model for Pr(X|$\boldsymbol{C}$). While other links can be applied, here we adopt the familiar logit link and assume that eqn. (6) is correctly specified:

$$logit[\Pr(X=1|C_1, C_2 \ldots C_Q)] = \gamma_0 + \sum_{q=1}^{Q} \gamma_q C_q. \quad (6)$$

The overall likelihood for the main study follows as:

$$L_m = \Pi_{i=1}^{n_m} \left\{ \sum_{y_i=0}^{y_i=1}\sum_{x_i=0}^{x_i=1} Pr\,(y^*{}_i|y_i)\, Pr\,(x^*{}_i|x_i) Pr\,(y_i|x_i, \boldsymbol{c_{iy}})\, Pr(x_i|\boldsymbol{c_{ix}}) \right\}, \quad (7)$$

where the subscripts of $\boldsymbol{C}$ in eqn.(7) indicate that the version of the $\boldsymbol{C}$ vector may differ across models as long as $\boldsymbol{c_{ix}}$ is a subset of $\boldsymbol{c_{iy}}$. The scenario just described assuming nondifferential misclassification of both X and Y is an extension of the classical 2×2 setting [2] to incorporate an arbitrary set of accurately measured covariates (**C**) into the model associating X with Y. In [2], Barron presented an identity, later termed the "matrix method", to connect the correct and error-prone cell counts (or probabilities) via the bridge of the diagnostic parameters SE_x, SE_y, SP_x and SP_y. Unless one is to rely upon sensitivity analyses, these parameters are best estimated by supplemental sampling of subjects into an internal validation set. According to the strategy discussed following eqn. (2), we assume that (X, X^*, Y, Y^*) are observed on each subject in the internal validation sample (although

only minor adjustments to the likelihood contributions are required, e.g., if certain subjects have only one variable validated). The likelihood contribution for the internal validation subsample is:

$$L_v=\prod_{j=1}^{n_v} Pr\left(y^*{}_i|y_i\right) Pr(x^*{}_i|x_i)\, Pr\left(y_i|x_i,\boldsymbol{c_{iy}}\right) Pr(x_i|\boldsymbol{c_{ix}}), \quad (8)$$

and the overall joint likelihood to be maximized is the product of the expressions in (7) and (8).

2.2.2 Independent differential misclassification—In contrast to the nondifferential case, differential misclassification occurs when the misclassification probabilities of one variable depend on the value(s) of the other variable(s). More specifically, regarding classifying via Y^*, we define

$$SE_{Yxc}=\Pr(Y^*=1|Y=1,X=x,\boldsymbol{C}=\boldsymbol{c})$$

and

$$SP_{Yxc}=\Pr(Y^*=0|Y=0,X=x,\boldsymbol{C}=\boldsymbol{c}). \quad (9)$$

Note that the SE and SP of Y now can be functions of exposure (X) and other covariates (***C***). For clarification, as long as the SE/SP's for X and Y depend only on the true values of other variables, we will term it "differential but independent misclassification". Eqn. (9) is the most general representation of SE_Y and SP_Y under such an assumption, and implies the possibility of modeling SE_Y and SP_Y parametrically. Similarly, X may or may not be an important factor in characterizing the diagnostic properties for Y, and can be left out when deemed not to be. We may define the misclassification process for X analogously, as follows:

$$SE_{Xyc}=\Pr(X^*=1|X=1,Y=y,\boldsymbol{C}=\boldsymbol{c})$$

and

$$SP_{Xyc}=\Pr(X^*=0|X=0,Y=y,\boldsymbol{C}=\boldsymbol{c}). \quad (10)$$

Each observation in the main study now contributes to the likelihood as follows:

$$\begin{aligned}\Pr(\mathrm{Y}^*=\mathrm{y}^*,\mathrm{X}^*=\mathrm{x}^*|\boldsymbol{C}=\boldsymbol{c})&=\sum_{\mathrm{y}=0}^{\mathrm{y}=1}\sum_{\mathrm{x}=0}^{\mathrm{x}=1}\Pr(\mathrm{y}^*,\mathrm{x}^*,\mathrm{Y},\mathrm{x}|\boldsymbol{c})\\ =\sum_{\mathrm{y}=0}^{\mathrm{y}=1}\sum_{\mathrm{x}=0}^{\mathrm{x}=1}&\Pr(\mathrm{y}^*|\mathrm{y},\mathrm{x},\boldsymbol{c})\Pr(\mathrm{x}^*|\mathrm{x},\mathrm{y},\boldsymbol{c})\Pr(\mathrm{y}|\mathrm{x},\boldsymbol{c})\Pr(\mathrm{x}|\boldsymbol{c}).\end{aligned} \quad (11)$$

Note that eqn. (11) is a generalization of (7) to incorporate differential misclassification. The first two terms in the last line of (11), characterizing SE and SP for X and Y, need to be specified in order to write out the likelihood. Here again we favor logistic regressions for modeling the misclassification processes of X and Y. More specifically, define

$$logit[\Pr(Y^*=1|Y,X,C_1,C_2\ldots C_R)]=\theta_0+\sum_{r=1}^{R}\theta_r C_r+\theta_{R+1}Y+\theta_{R+2}X \quad (12)$$

and

$$logit[\Pr(X^*=1|Y,X,C_1,C_2\ldots C_S)]=\delta_0+\sum_{s=1}^{S}\delta_r C_r+\delta_{S+1}X+\delta_{S+2}Y. \quad (13)$$

Model (12) implies that

$$SE_{Yxc}=\frac{\exp(\theta_0+\sum_{r=1}^{R}\theta_r C_r+\theta_{R+1}+\theta_{R+2}X)}{1+\exp(\theta_0+\sum_{r=1}^{R}\theta_r C_r+\theta_{R+1}+\theta_{R+2}X)}$$

and

$$SP_{Yxc}=\frac{1}{1+\exp(\theta_0+\sum_{r=1}^{R}\theta_r C_r+\theta_{R+2}X)}$$

The corresponding definitions of SE_{Xyc} and SP_{Xyc} are determined analogously from model (13). The full likelihood follows as $L=L_m\times L_v$, where

$$L_m=\prod_{i=1}^{n_m}\left\{\sum_{y_i=0}^{y_i=1}\sum_{x_i=0}^{x_i=1}Pr\,(y^*_i|y_i,x_i,\boldsymbol{c_{iy*}})\,Pr\,(x^*_i|x_i,y_i,\boldsymbol{c_{ix*}})Pr\,(y_i|x_i,\boldsymbol{c_{iy}})\,Pr(x_i|\boldsymbol{c_{ix}})\right\} \quad (14)$$

and

$$L_v=\prod_{j=1}^{n_v}Pr\,\left(y^*_j|y_j,x_j,\boldsymbol{c_{jy*}}\right)Pr(x^*_j|x_j,y_j,\boldsymbol{c_{jx*}})Pr\,(y_j|x_j,\boldsymbol{c_{jy}})\,Pr(x_j|\boldsymbol{c_{jx}}) \quad (15)$$

The major distinction of eqn.s. (14) and (15) from eqn.s. (7) and (8) is the incorporation of differential misclassification via the modeling of the SE and SP parameters based on (12) and (13). Nondifferential misclassification may be viewed as a special case, corresponding to the assumption that ($\theta_1=\ldots=\theta_R=\theta_{R+2}=0$) in (12), or that ($\delta_1=\ldots=\delta_S=\delta_{S+2}=0$) in (13). The Akaike information criterion (AIC) [31] approach based on the internal validation study design offers the distinct advantage of helping to assess whether nondifferentiality best describes the data, as well as aiding in model selection for screening out factors associated with SE and SP for both X and Y. The scenario considered in this section is a generalization

of that considered in [26], by allowing covariates in the main model and in the models for misclassification.

2.2.3 Dependent and differential misclassification—In the previous section, the diagnostic properties (SE and SP) of X and Y were allowed to be impacted by the true values of other variables, allowing flexible models for differential misclassification. With two binary variables subject to error, one must also recognize the potential for "dependent" misclassification (e.g., [10]), where the SE and SP for one variable may be conditionally dependent on the error-prone assessment of the other. To clarify this notion of dependence in our setting, consider the following main study likelihood contribution:

$$Pr(Y^*=y^*, X^*=x^*|\boldsymbol{C}=\boldsymbol{c})=\sum_{y=0}^{y=1}\sum_{x=0}^{x=1} Pr(y^*, x^*, y, x|\boldsymbol{c})= \sum_{y=0}^{y=1}\sum_{x=0}^{x=1} Pr(y^*|y, x, x^*, \boldsymbol{c})Pr\,(x^*|x, y, \boldsymbol{c})Pr(y|x, \boldsymbol{c})Pr(x|\boldsymbol{c}). \quad (16)$$

The expression in (16) reflects a model that generalizes (6) and (11) to encompass both dependence and differentiality. The first term in the last line of eqn. (16) reflects dependent misclassification by conditioning on X^*, i.e., generalizing the SE and SP for Y* to depend on X* conditionally on (Y, X, and **C**). The full main/internal validation study likelihood under such a model remains $L=L_m \times L_v$, where now the likelihood contribution for a main study subject is

$$\sum_{y_i=0}^{y_i=1}\sum_{x_i=0}^{x_i=1} Pr\,({y^*}_i|y_i, x_i, {x_i}^*, \boldsymbol{c_{iy*}})\, Pr({x^*}_i|x_i, y_i, \boldsymbol{c_{ix*}})Pr\,(y_i|x_i, \boldsymbol{c_{iy}})\, Pr(x_i|\boldsymbol{c_{ix}}).$$

Each internal validation subject now makes the following likelihood contribution:

$$Pr\,\left({y^*}_j|y_j, x_j, {x_j}^*, \boldsymbol{c_{jy*}}\right)\, Pr({x^*}_j|x_j, y_j, \boldsymbol{c_{jx*}})Pr(y_j|x_j, \boldsymbol{c_{jy}})Pr(x_j|\boldsymbol{c_{jx}}).$$

As before, proper model selection may make it possible to reduce the model to a simpler form. For example, one can rely on AIC to assess the appropriateness of independent misclassification in (16), using the following generalized logistic regression model for the error-prone binary Y*:

$$logit[\Pr(Y^*=1|Y, X, X^*, C_1, C_2 \ldots C_r)]=\theta_0+\sum_{i=1}^{R}\theta_r\, C_r+\theta_{R+1}Y+\theta_{R+2}X+\theta_{R+3}X^*. \quad (17)$$

The AIC can be used to address whether to include θ_{R+3}.

2.2.4 Other types of misclassification—It should be noted that there are other possible misclassification mechanisms besides those discussed in the previous sections, which may be regarded as special cases of the model presented in Section 2.2.3. For example, in practice, both exposure and disease status could be nondifferentially misclassified, yet misclassification could be dependent [31]. In such a case, one may make the appropriate simple adjustments to the form of the likelihood. Other mechanisms include the possibility

that disease and exposure status are independently misclassified, with misclassification differential for one and nondifferential for the other. All likelihood functions presented herein for the main/internal validation study design potentially permit convenient numerical maximization via general optimization facilities available in standard statistical software such as SAS NLMIXED. Details about the optimization techniques available in NLMIXED can be found in SAS documentation [32]. Standard errors for estimates of secondary parameters (e.g., SE and SP) can be estimated via the delta method, which is also part of the routine procedure in NLMIXED. Readily adaptable SAS code for this purpose is attached in the Appendix, using the HERS data in Section 3 as an example. In practice, thorough model selection should be performed for the SE/SP models for X and Y using the joint main/validation sample, to assess whether dependence and/or differentiality is involved in the misclassification process.

2.3 Factorizing the general misclassification model

Note the following alternative factorization for the general likelihood contribution in (16):

$$Pr(Y^*=y^*, X^*=x^*|\boldsymbol{C}=\boldsymbol{c})=\sum_{y=0}^{y=1}\sum_{x=0}^{x=1} Pr(y^*, x^*, y, x|\boldsymbol{c})$$
$$=\sum_{y=0}^{y=1}\sum_{x=0}^{x=1} Pr(x^*|x, y, y^*, \boldsymbol{c})Pr(y^*|y, x, \boldsymbol{c})Pr(y|x, \boldsymbol{c})Pr(x|\boldsymbol{c})$$

The last two terms in the final expression above are identical to those in eqn. (16), although the misclassification models for X and Y are altered in the dependent and differential misclassification case. However, this alternative factorization has no effect on the specifications in Section 2.2, under less than fully general misclassification patterns. In practice, one can use AIC to select the best-fitting factorization for the misclassification mechanism in the general case.

2.4 Variable selection for X|C and misclassification models

In principle, the covariates in the X|**C** or misclassification models could differ from those in the primary response model of interest. However, caution should be applied in this regard. Consider a situation in which a variable, say V, is included in the X|**C** model, and Y is not independent on V conditional on X and all other covariates. In that case, failing to include V in the response model will generally induce bias in the estimation. A similar argument holds for variables involved in misclassification models. All variables in misclassification models should be accounted for in the response model, unless there is evidence suggesting that it is legitimate to assume they are independent of Y conditional on X and other covariates in the response model. In Sections 4.2, we have provided empirical evidence for the above argument. In practice, we make the following suggestion. First, the response model of interest can be selected based upon the scientific research question and preliminary model selection to screen out the maximal ***C*** vector for all models, including the X|**C** and misclassification models. In other words, only a subset of ***C*** will be considered in the X|**C** and misclassification models. Then an AIC-based model selection strategy can be conducted by fitting candidate models and choosing the one with the smallest AIC.

3. HERS example analysis

According to eqn (1), age, race, risk cohort and HIV status are considered candidate ***C*** covariates for all models. In order to demonstrate the performance of the proposed approach, we randomly selected 1/4 of the total HERS sample size at visit 4 (n_v=214) into an internal validation subsample. Model selection on all 214 participants suggested a version of the X|C model as follows:

$$logit[\Pr(CULTURE\ TRICH=1)]=\gamma_0+\gamma_1 RACE \quad (18)$$

where race is associated with Trichomoniasis status based on the gold standard culture method. Predictor selection based on those 214 women further suggested dependent and differential misclassification of the CLIN BV and WET TRICH methods. The selected SE/SP models for CLIN BV and WET TRICH are as follows:

$$logit[\Pr(CLINBV=1)]=\theta_0+\theta_1 LABBV+\theta_2 WETTRICH+\theta_3 CULTURETRICH+\theta_4 RISKCHRT+\theta_5 HIVPOS \quad (19)$$

$$logit[\Pr(WETTRICH=1)]=\delta_0+\delta_1 CULTURETRICH+\delta_2 LABBV+\delta_3 RISKCHRT \quad (20)$$

Thus, the classification rates of CLIN BV are found to be dependent on the risk factors of HIV risk cohort and HIV status, implying differential misclassification. Conditionally on these variables, the misclassification process for BV depends on the error-prone wet mount version of the Trichomoniasis diagnosis along with the culture Trichomoniasis diagnosis, implying dependent misclassification. Similarly, model (20) indicates that HIV risk cohort significantly impacts the SE and SP of the wet mount method, a typical example of differential misclassification. As a practical matter, it is useful to perform model selection on the internal validation sample for eqns. (18)–(20) via a traditional likelihood ratio testing approach with a conservative significance threshold such as 0.10 or 0.15, to ensure important factors are not left out. As recommended in Section 2.4, AIC can then be applied to the joint model to compare candidate misclassification mechanisms.

The first model fitted in Table 2 reflects the joint modeling of eqns. (1) and (18)–(20) via the likelihood contributions given in (16) for dependent and differential misclassification. Note that this analysis incorporating gold standard test results (Y, X) for only a random subsample of women yields the same interpretation as the ideal analysis (Table 1), with all primary parameter estimates having similar magnitudes and the same directionalities. In contrast, results assuming independent differential or nondifferential misclassification are much more similar to the results of the naïve analysis. For instance, when assuming independence but allowing differential misclassification, the estimate for Trichomoniasis status increases, with a similar magnitude as observed in the naïve analysis. If assuming independent and nondifferential misclassification, a highly elevated estimate for HIV risk cohort and a negative estimated ln(OR) for HIV status are obtained, in addition to the elevated estimate for Trich status. Note that AIC favors the most general misclassification model, suggesting a need to account for both dependent and differential misclassification in the HERS example. This important observation clearly highlights the value of internal

validation sampling for evaluating and modeling potentially complex misclassification mechanisms. We also note that the alternative factorization (Section 2.7) of the dependent differential misclassification model was assessed, but was not selected due to an AIC value larger than that for the factorization in Table 2.

Tables 3–4 summarize the ML estimates for misclassification parameters charactering CLIN BV and wet mount Trichomoniasis in different strata, based on the selected models in (19) and (20). Note that Trich diagnosis, HIV risk cohort, and HIV status all significantly affect the diagnostic properties of CLIN BV. By holding other covariates constant, SE tends to be higher in wet mount or culture Trichomoniasis positive patients and intravenous drug users, while lower in HIV positive women. An opposite trend is observed for the SP estimates. For wet mount Trichomoniasis, intravenous drug users or culture Trichomoniasis positive participants appear to have a greater SE than those at risk via sexual contact, while the test is highly specific in both groups.

One might wonder whether fitting the full joint likelihood using the complete set of error-prone and gold standard test results for BV and Trichomoniasis would yield different results than those based only on the gold standard data. The wide availability of the four variables (X, Y, X^*, Y^*) in the HERS permits this unique comparison, and we find unsurprisingly that the results are exactly the same as for the "Ideal Analysis" in Table 1. The reason is that with (X, Y, X^*, Y^*) available on all subjects, the likelihood factors into four separate models (eqns. (1) and (18)–(20)). In that case, estimates of the parameters in eqn. (1) do not rely on estimates from the other models.

4. Simulation studies

4.1 Data-driven simulations

The simulation experiment summarized in Table 5 demonstrates the empirical performance of estimation based on the joint main/internal validation likelihood, under dependent differential misclassification conditions mimicking those observed at HERS visit 4. Specifically, data were generated jointly under the primary model of interest (1) and the logistic models in (18)–(20), with true coefficients and covariate distributions similar to those estimated and observed in the HERS example. The ideal and naïve analyses (as in Table 1) were based on 500 simulated replicates of (X,Y) and (X^*, Y^*) respectively, while the complete analysis was conducted via joint ML using the main and internal validation data corresponding to each replication as described in Section 2. The internal validation subsample was randomly selected from 25%, 10% and 5% of the total sample, while the total sample size was 904, as in the HERS example.

Table 5 confirms that the naïve analysis yields markedly biased estimates, with correspondingly poor confidence interval (CI) coverage. The joint likelihood-based analysis assuming dependent and differential misclassification, however, produces reliable results with mean coefficient estimates very close to the true values and excellent 95% CI coverage, given a sufficient internal validation sample. Empirical evidence in this setting mimicking the HERS suggests that an internal validation sample of approximately 15% of the total sample size (i.e., 226), is sufficient to ensure valid parameter estimates corrected for

complex misclassification. As the internal sample size drops to 10% (around 90 subjects) or 5% (around 45 subjects), the corrected estimate of the exposure coefficient in particular begins to suffer bias. Predictably, inflated standard errors are also observed with smaller internal validation samples. In practice, we thus recommend empirical examinations at the study planning stage based upon a series of assumptions about the misclassification processes. Simulation studies using programs available from the lead author can facilitate such examinations to inform the necessary total sample size and validation fraction.

4.2 Impact of X|C model misspecification

Simulation studies were also conducted to examine the impact of misspecifying the X|**C** model on estimating primary parameters. In hypothetical examples, data were simulated based on eqns. (1) and (18)–(20) with HIV status an important predictor in both the primary Y|X,**C** model and the X|**C** model. A total of 500 simulations were performed in Table 6, which also provides the true parameter values and overall and internal validation sample sizes. Three candidate models were fit under each scenario: the correct model, a model misspecifying the X|**C** model and a model misspecifying the primary model.

Table 6 shows that when HIV status is important in both the X|**C** and primary models, failing to adjust for this factor in the X|**C** model appears to slightly bias the estimate of the coefficient of HIV status in the primary model. When the X|**C** model is correctly specified but HIV status is omitted from the primary model, the estimated coefficients for TRICH and race in the latter are biased. We also simulated the scenario when HIV status is only important in the X|**C** model. In this case we observe that the accuracy in estimating primary parameters remains (results not shown). Similar findings have been observed in simulations when misclassification models are subject to misspecification. Empirical observations support adopting the strategy recommended in Section 2.4, i.e., that one should restrict the attention only to covariates selected in the primary model when building the X|**C** and misclassification models. Perhaps most importantly, we find that the AIC-based model selection strategy selected the correct model in most cases.

Discussion

In this manuscript, we have expanded upon previously demonstrated likelihood-based methodology for dealing with misclassification via a main/internal validation study design. The approach may be viewed as a generalization of aspects of prior work [e.g., 2, 26] to incorporate validation data and covariates into both the main model of interest and into misclassification models. It also generalizes the logistic regression-based approach outlined in [24], in which only outcome misclassification was addressed. The joint likelihood is developed for various combinations of differential and/or dependent misclassification of two binary variables (Y and X), yielding special case extensions of general main/internal validation study likelihood specifications described in [14]. Specific motivation for these developments is provided by cross-sectional data involving assessments of BV and Trichomoniasis status and covariates measured in the HERS.

Our goal has been to provide clear guidance on adjusting for biases due to misclassification in binary response and predictor variables in logistic regression, when resources permit the

collection of a reasonably-sized internal sample with validations of both variables. We strongly emphasize the importance of such internal validation sampling to assess misclassification patterns and to ensure the validity of the results, noting that only this study design-based tactic permitted the identification of the complex misclassification mechanisms operating in the HERS. The parametric approach taken here has the advantage of accessibility, with SAS NLMIXED programs provided in the Appendix. It also makes AIC calculations available to assist with model selection and mechanism evaluation, as demonstrated in the HERS example.

Although throughout we have adopted logit links for all models, other links can also be used without noteworthy additional conceptual or technical difficulty. Compared to alternative approaches such as sensitivity analyses, our approach along with the collection of internal validation data allows much greater flexibility when the misclassification process is complex, as evident in the HERS data with respect to BV and Trichomoniasis status. In contrast, sensitivity analyses may be preferred if there is a strong reason to believe that misclassification in both variables is nondifferential and if there is reasonably reliable preliminary knowledge regarding the operating characteristics (SE and SP) of the error-prone tests measuring the response and the exposure. However, if misclassification may be complex, it would be valuable to make efforts to collect validation information.

As with any parametric approach, model selection can be critical in our setting in order to ensure valid estimation. We have suggested a practical strategy in Section 2.4, in which candidate covariates in the X|**C** and SE/SP models should only be selected from covariates in the primary model, unless one has confidence that additional covariates introduced into the X|**C** or misclassification models are conditionally independent from the response. Otherwise, bias may occur in the estimation. An alternative analytic strategy could borrow ideas from multiple imputation [34], via which one could base imputation of the value of X on a model that includes other correctly measured covariates (***C***). Cole *et al.* [35] demonstrate the application of multiple imputation when X is misclassified. Although only nondifferential misclassification was discussed in their work, their approach could potentially be extended to more general situations that might include differential misclassification and the case of both X and Y subject to misclassification. Compared to existing alternatives, we note that our ML approach has been generalized to complex misclassification in both exposure and response variables, and is computationally accessible. It also allows for AIC-based model selection, which makes it possible to carefully study and account for the misclassification pattern. One should always note that, like in all model selection problems, the primary response model should be specified in light of the research question as well as scientific knowledge, in addition to statistical considerations.

In separate work [36], we have studied methods to adjust for differential misclassification of BV status longitudinally within the HERS. Future work may include efforts to extend these regression-based correction approaches to adjust for both outcome and predictor misclassification when both are repeatedly measured over time.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by grants from the National Institute of Nursing Research (1RC4NR012527-01), the National Institute of Environmental Health Sciences (5R01ES012458-07), and the National Center for Advancing Translational Sciences (UL1TR000454). The HIV Epidemiology Research Study (HERS) was supported by the Centers for Disease Control and Prevention (CDC): U64/CCU106795, U64/CCU206798, U64/CCU306802, U64/CCU506831. The content is solely the findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the National Institutes of Health or the Centers for Disease Control and Prevention.

The authors especially thank the HERS participants and the HERS Group, which consists of: Robert S. Klein, M.D., Ellie Schoenbaum, M.D., Julia Arnsten, M.D., M.P.H., Robert D. Burk, M.D., Chee Jen Chang, Ph.D., Penelope Demas, Ph.D., and Andrea Howard, M.D., M.Sc., from Montefiore Medical Center and the Albert Einstein College of Medicine; Paula Schuman, M.D. and Jack Sobel, M.D., from the Wayne State University School of Medicine; Anne Rompalo, M.D., David Vlahov, Ph.D. and David Celentano, Ph.D., from the Johns Hopkins University School of Medicine; Charles Carpenter, M.D. and Kenneth Mayer, M.D. from the Brown University School of Medicine; Ann Duerr, M.D., Caroline C. King, Ph.D., Lytt I. Gardner, Ph.D., Charles M. Heilig, PhD., Scott Holmberg, M.D., Denise Jamieson, M.D., Jan Moore, Ph.D., Ruby Phelps, B.S., Dawn Smith, M.D., and Dora Warren, Ph.D. from the CDC; and Katherine Davenny, Ph.D. from the National Institute of Drug Abuse.

References

1. Bross IDJ. Misclassification in 2×2 tables. Biometrics. 1954; 10:478–486.
2. Barron BA. The effects of misclassification on the estimation of relative risk. Biometrics. 1977; 33:414–417. [PubMed: 884199]
3. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. American Journal of Epidemiology. 1977; 105:488–495. [PubMed: 871121]
4. Greenland S. The effect of misclassification in the presence of covariates. American Journal of Epidemiology. 1980; 112:564–569. [PubMed: 7424903]
5. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? American Journal of Epidemiology. 1990; 132:746–748. [PubMed: 2403115]
6. Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. Epidemiology. 1992; 3:210–215. [PubMed: 1591319]
7. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease. Relationships and methods of correction. Annual Review of Public Health. 1993; 14:69–93.
8. Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. Biometrika. 1999; 86:843–855.
9. Green MS. Use of predictive value to adjust relative risk estimates biases by misclassification of outcome status. American Journal of Epidemiology. 1983; 117:98–105. [PubMed: 6823958]
10. Greenland S, Kleinbaum DG. Correcting for misclassification in two way tables and matched-pair studies. International Journal of Epidemiology. 1983; 12:93–97. [PubMed: 6840961]
11. Greenland S. Variance estimation of epidemiologic effect estimates under misclassification. Statistics in Medicine. 1988; 7:745–757. [PubMed: 3043623]
12. Pepe MS. Inference using surrogate outcome data and a validation sample. Biometrika. 1992; 79:355–365.
13. Brenner H, Gefeller O. Use of positive predictive value to correct for diease misclassification in epidemiologic studies. American Journal of Epidemiology. 1993; 138:1007–1015. [PubMed: 8256775]

14. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement Error in Nonlinear Models. 2. London: Chapman and Hall; 2006.
15. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. American Journal of Epidemiology. 1997; 146:195–203. [PubMed: 9230782]
16. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology. 2003; 14:451–458. [PubMed: 12843771]
17. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. International Journal of Epidemiology. 2005; 34:1370–1376. [PubMed: 16172102]
18. Lyles RH, Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. Statistics in Medicine. 2010; 29:2297–2309. [PubMed: 20552681]
19. Marshall RJ. Validation study methods for estimating proportions and odds ratios with misclassified data. Journal of Clinical Epidemiology. 1990; 43:941–947. [PubMed: 2213082]
20. Morrissey MJ, Spiegelman D. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. Biometrics. 1999; 55:338–344. [PubMed: 11318185]
21. Lyles RH. A note on estimating crude odds ratios in case–control studies with differentially misclassified exposure. Biometrics. 2002; 58:1034–1037. [PubMed: 12495160]
22. Greenland S. Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. Journal of Statistical Planning and Inference. 2008; 138:528–538.
23. Smith DK, Warren DL, Vlahov D, Schuman P, Stein MD, Greenberg BL. Design and baseline participant characteristics of the Human Immunodeficiency Virus Epidemiology Research (HER) Study: A prospective cohort study of human immunodeficiency virus infection in U.S. women. American Journal of Epidemiology. 1997; 146:459–469. [PubMed: 9290506]
24. Lyles RH, Tang L, Superak HM, King CC, Celantano D, Lo Y, Sobel J. An illustration of validation data-based adjustments for outcome misclassification in logistic regression. Epidemiology. 2011; 22:589–597. [PubMed: 21487295]
25. Buonaccorsi, J. Measurement Error: Models, Methods and Applications. Chapman Hall; 2010.
26. Tang L, Lyles RH, Ye Y, Lo Y, King CC. Extending “Matrix” and “Inverse Matrix” Methods: Another Look at Barron’s Approach. Epidemiologic Methods. 2013; 2:49–66. [PubMed: 25844304]
27. Demirezen S, Korkmaz E, Beksac MS. Association between Trichomoniasis and bacterial vaginosis: examination of 600 cervicovaginal smears. Central European Journal of Public Health. 2005; 13:96–98. [PubMed: 15969458]
28. Amsel R, Totten PA, Spiegel CA, Chen KC, Eschenbach D, Holmes KK. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. American Journal of Medicine. 1983; 74:14–22. [PubMed: 6600371]
29. Thomason JL, Gelbart SM, Sobun JF, Schulien MB, Hamilton PR. Comparison of four methods to detect Trichomonas vaginalis. Journal of Clinical Microbiology. 1988; 26:1869–1870. [PubMed: 3141470]
30. Nugent RP, Krohn MA, Hillier SL. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. Journal of Clinical Microbiology. 1991; 29:297–301. [PubMed: 1706728]
31. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19:716–723.
32. Rothman, KJ.; Greenland, S.; Lash, TL. Modern Epidemiology. 3. Lippincott Williams & Wilkins; 2008.
33. SAS Institute, Inc. SAS/STAT 9.1 User’s Guide. Cary, NC: SAS Institute, Inc; 2004.
34. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley & Sons; 1987.
35. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. International Journal of Epidemiology. 2006; 35:1074–1081. [PubMed: 16709616]
36. Tang L, Lyles RH, King CC, Hogan JW, Lo Y. Regression analysis for differentially misclassified correlated binary outcomes. Journal of the Royal Statistical Society, Series C. 2015

Table 1

Logistic Regression Results for BV Status (Y) on 904 Women at the 4th HERS Visit.

Variable	$\hat{\beta}$ (StdErr)	$\hat{OR}$ (95% CI)
Ideal Analysis[a]		
Trichomoniasis (+ vs. −)	0.88 (0.19)	2.41 (1.66, 3.50)
Age (Years)	−0.04 (0.01)	0.96 (0.94, 0.98)
Race (Black vs. Others)	0.76 (0.16)	2.15 (1.57, 2.92)
HIV Risk Cohort (IDU vs. Sexual)	0.31 (0.15)	1.37 (1.03, 1.83)
HIV Status (+ vs. −)	0.22 (0.15)	1.25 (0.93, 1.69)
Naive Analysis[b]		
Trichomoniasis (+ vs. −)	1.24 (0.27)	3.44 (2.03, 5.84)
Age (Years)	−0.05 (0.01)	0.95 (0.93, 0.98)
Race (Black vs. Others)	0.69 (0.18)	1.99 (1.40, 2.83)
HIV Risk Cohort (IDU vs. Sexual)	0.90 (0.17)	2.45 (1.74, 3.45)
HIV Status (+ vs. −)	−0.31 (0.17)	0.73 (0.52, 1.03)

[a] LAB BV vs Culture Trichomoniasis, adjusting for age, race, HIV risk cohort and HIV status.

[b] CLIN BV vs Wet Mount Trichomoniasis, adjusting for age, race, HIV risk cohort and HIV status.

Table 2

Results of ML Analysis of Main/Internal Validation Study Data on 904 Women at the 4th HERS Visit (n_m=690, n_v=214).

Variable	$\hat{\beta}$ (StdErr)	$\hat{OR}$ (95% CI)
Assuming dependent and differential misclassification[a] (AIC=1771.5)		
Trichomoniasis (+ vs. −)	0.64 (0.41)	1.90 (0.38, 3.42)
Age (Years)	−0.05 (0.02)	0.95 (0.92, 0.98)
Race (Black vs. Others)	0.79 (0.24)	2.19 (1.16, 3.23)
HIV Risk Cohort (IDU vs. Sexual)	0.28 (0.26)	1.32 (0.64, 2.00)
HIV Status (+ vs. −)	0.23 (0.27)	1.26 (0.58, 1.94)
Assuming independent and differential misclassification[b] (AIC=1777.7)		
Trichomoniasis (+ vs. −)	1.06 (0.38)	2.88 (0.73, 5.02)
Age (Years)	−0.05 (0.02)	0.95 (0.92, 0.98)
Race (Black vs. Others)	0.73 (0.24)	2.07 (1.09, 3.05)
HIV Risk Cohort (IDU vs. Sexual)	0.30 (0.26)	1.35 (0.65, 2.05)
HIV Status (+ vs. −)	0.21 (0.28)	1.23 (0.56, 1.90)
Assuming nondifferential misclassification[c] (AIC=1809.8)		
Trichomoniasis (+ vs. −)	1.38 (0.37)	3.99 (1.09, 6.90)
Age (Years)	−0.05 (0.02)	0.95 (0.92, 0.99)
Race (Black vs. Others)	0.73 (0.24)	2.08 (1.11, 3.04)
HIV Risk Cohort (IDU vs. Sexual)	0.85 (0.22)	2.35 (1.32, 3.37)
HIV Status (+ vs. −)	−0.14 (0.23)	0.87 (0.47, 1.26)

[a] ML estimates of primary parameters are obtained by jointly modeling eqn.s (1) and (18)–(20).

[b] WETTRICH is removed from eqn. (19) to indicate independence. The assumption is not supported by the data from the AIC value.

[c] No covariates affect SE and SP of Y and X in eqn.s (19) and (20). The assumption is not supported by the data from the AIC value.

Table 3

Results of ML Analysis of Main/Internal Validation Study Data on 904 Women at the 4th HERS Visit (n_m=690, n_v=214): Estimates of eqn. (19)[a] Describing SE/SP of Y^* (CLIN BV).

ML estimates of coefficients in eqn. (19)		
Variable	$\widehat{\theta}$**(95% CI)**	**p-value**
LAB BV (+ vs. −)	2.92 (1.99, 3.85)	<0.0001
Wet Mount Trichomoniasis (+ vs. −)	1.34 (0.33, 2.34)	0.01
Culture Trichomoniasis (+ vs. −)	0.36 (−0.21, 0.94)	0.22
HIV Risk Cohort (IDU vs. Sexual)	0.91 (0.41, 1.41)	0.0004
HIV Status (+ vs. −)	−0.61 (−1.15, −0.06)	0.03

[a] ML estimates of parameters in eqn (19) are obtained by jointly modeling eqns (1) and (18)–(20).

Table 4

Results of ML Analysis of Main/Internal Validation Study Data on 904 Women at the 4th HERS Visit (n_m=690, n_v=214): Estimates of eqn. (20)[a] Describing SE/SP of X* (WET TRICH).

ML estimates of coefficients in eqn. (20)		
Variable	$\hat{\delta}$**(95% CI)**	**p-value**
Culture Trichomoniasis (+ vs. −)	4.13 (2.39, 5.87)	<0.0001
LAB BV (+ vs. −)	0.45 (−0.26, 1.16)	0.21
HIV Risk Cohort (IDU vs. Sexual)	1.28 (0.55, 2.00)	0.001

[a] ML estimates of parameters in eqn (20) are obtained by jointly modeling eqns (1) and (18)–(20).

Table 5

Results of Simulations Designed to Mimic Conditions of HERS Example[a].

Variable	$\hat{\beta}$(StdErr)[h]	95% CI Coverage
Ideal Analysis[b]		
Trichomoniasis (+ vs. −)	0.65 (0.25)	96.6%
Age (Years)	−0.05 (0.003)	93.2%
Race (Black vs. Others)	0.80 (0.19)	96.4%
HIV Risk Cohort (IDU vs. Sexual)	0.28 (0.18)	95.0%
HIV Status (+ vs. −)	0.25 (0.19)	96.4%
Naïve Analysis[c]		
Trichomoniasis (+ vs. −)	1.54 (0.29)	12.2%
Age (Years)	−0.02 (0.002)	0
Race (Black vs. Others)	0.38 (0.18)	37.2%
HIV Risk Cohort (IDU vs. Sexual)	0.83 (0.18)	12.8%
HIV Status (+ vs. −)	−0.42 (0.18)	5.8%
Complete Analysis (n_v=25%×n)[d]		
Trichomoniasis (+ vs. −)	0.61 (0.49)	94.6%
Age (Years)	−0.05 (0.01)	96.0%
Race (Black vs. Others)	0.82 (0.34)	95.0%
HIV Risk Cohort (IDU vs. Sexual)	0.28 (0.33)	94.6%
HIV Status (+ vs. −)	0.23 (0.35)	95.6%
Complete Analysis (n_v=15%×n)[e]		
Trichomoniasis (+ vs. −)	0.58 (0.62)	94.2%
Age (Years)	−0.05 (0.01)	93.6%
Race (Black vs. Others)	0.82 (0.40)	94.2%
HIV Risk Cohort (IDU vs. Sexual)	0.28 (0.36)	95.0%
HIV Status (+ vs. −)	0.24 (0.42)	95.8%
Complete Analysis (n_v=10%×n)[f]		
Trichomoniasis (+ vs. −)	0.50 (0.76)	93.0%
Age (Years)	−0.05 (0.01)	93.6%
Race (Black vs. Others)	0.84 (0.45)	94.0%
HIV Risk Cohort (IDU vs. Sexual)	0.28 (0.48)	93.6%
HIV Status (+ vs. −)	0.26 (0.50)	93.6%
Complete Analysis (n_v=5%×n)[g]		
Trichomoniasis (+ vs. −)	0.23 (1.05)	84.6%
Age (Years)	−0.05 (0.02)	93.0%
Race (Black vs. Others)	0.94 (0.58)	94.4%
HIV Risk Cohort (IDU vs. Sexual)	0.35 (0.62)	95.2%

Variable	$\hat{\beta}$(StdErr)[h]	95% CI Coverage
HIV Status (+ vs. −)	0.29 (0.65)	93.4%

[a] 500 simulations. n_m=690, n_v=214. Data were generated from eqns. (1) and (18)–(20). True parameters: (β_0=0.63, β_1=0.64, β_2=−0.05, β_3=0.79, β_4=0.28, β_5=0.23, θ_0=−3.27, θ_1=2.92, θ_2=1.34, θ_3=0.36, θ_4=0.91, θ_5=−0.61, γ_0=−3.43, γ_1=2.48, δ_0=−5.64, δ_1=4.13, δ_2=0.45, δ_3=1.28).

[b] ML estimates from eqn. (1).

[c] ML estimates from eqn. (1) with (Y^*, X^*) replacing (Y, X).

[d] ML estimates of primary parameters are obtained by jointly modeling eqn.s. (1) and (18)–(20). The internal validation sample accounts for 25% of the total sample. 500/500 replicates converged.

[e] The internal validation sample accounts for 15% of the total sample. 499/500 replicates converged.

[f] The internal validation sample accounts for 10% of the total sample. 497/500 replicates converged.

[g] The internal validation sample accounts for 5% of the total sample. 485/500 replicates converged.

[h] Empirical standard deviations across 500 estimates are reported in parenthesis.

Table 6

Results of Examining Impact of Misspecifying X|**C** Model When HIV Status is Important in X|**C** and Conditionally Dependent on Response[a].

Variable	$\hat{\beta}$(StdErr)[f]	
Correct model[b]		
Trichomoniasis (+ vs. −)	0.60 (0.44)	
Age (Years)	−0.05 (0.01)	
Race (Black vs. Others)	0.82 (0.35)	
HIV Risk Cohort (IDU vs. Sexual)	0.28 (0.32)	
HIV Status (+ vs. −)	1.05 (0.33)	
X|C is misspecified[c]		
Trichomoniasis (+ vs. −)	0.58 (0.44)	
Age (Years)	−0.05 (0.01)	
Race (Black vs. Others)	0.82 (0.33)	
HIV Risk Cohort (IDU vs. Sexual)	0.27 (0.32)	
HIV Status (+ vs. −)	1.09 (0.33)	
Primary Model is Misspecified[d]		
Trichomoniasis (+ vs. −)	0.88 (0.45)	
Age (Years)	−0.05 (0.01)	
Race (Black vs. Others)	0.65 (0.32)	
HIV Risk Cohort (IDU vs. Sexual)	0.26 (0.33)	
HIV Status (+ vs. −)	N/A	
ML estimates from AIC-based selected model[e]		
Variable	$\hat{\beta}$(StdErr)[f]	**95% CI Coverage**
Trichomoniasis (+ vs. −)	0.60 (0.44)	94.8%
Age (Years)	−0.05 (0.01)	95.2%
Race (Black vs. Others)	0.82 (0.34)	96.8%
HIV Risk Cohort (IDU vs. Sexual)	0.28 (0.32)	94.0%
HIV Status (+ vs. −)	1.07 (0.33)	95.2%

[a] 500 simulations. n_m=690, n_v=214. Data were simulated from eqns. (1) and (18)–(20). True parameters: (β_0=0.63, β_1=0.64, β_2=−0.05, β_3=0.79, β_4=0.28, β_5=1, θ_0=−3.27, θ_1=2.92, θ_2=1.34, θ_3=0.36, θ_4=0.91, γ_0=−3.43, γ_1=2.48, γ_2=1 for HIV status, δ_0=−5.64, δ_1=4.13, δ_2=0.45, δ_3=1.28).

[b] ML estimates of primary parameters are obtained by jointly modeling eqns. (1) and (18)–(20) with correct model specification.

[c] ML estimates of primary parameters are obtained by jointly modeling eqns. (1) and (18)–(20) with eqn. (18) misspecified in which HIV status was omitted.

[d] ML estimates of primary parameters are obtained by jointly modeling eqns. (1) and (18)–(20) with eqn. (1) misspecified in which HIV status was omitted.

[e] The model specifying both X|C and primary models correctly was selected 96% of the time.

[f] Empirical standard deviations across 500 estimates are reported in parenthesis.