Published in final edited form as:

J R Stat Soc Ser C Appl Stat. 2015 April ; 64(3): 433–449. doi:10.1111/rssc.12081.

Regression Analysis for Differentially Misclassified Correlated Binary Outcomes

Li Tang^{1,*}, Robert H. Lyles², Caroline C. King³, Joseph W. Hogan⁴, and Yungtai Lo⁵

¹ Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, U.S.A

² Department of Biostatistics and Bioinformatics, Rollins School of Public Health of Emory University, Atlanta, Georgia 30322, U.S.A

³ Division of Reproductive Health, Centers for Disease Control and Prevention, Atlanta, Georgia 30322, U.S.A

⁴ Center for Statistical Sciences, Program in Public Health, Brown University, Providence, RI 02912, U.S.A

⁵ Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, U.S.A

Summary

In many epidemiological and clinical studies, misclassification may arise in one or several variables, resulting in potentially invalid analytic results (e.g., estimates of odds ratios of interest) when no correction is made. Here we consider the situation in which correlated binary response variables are subject to misclassification. Building upon prior work, we provide an approach to adjust for potentially complex differential misclassification via internal validation sampling applied at multiple study time points. We seek to estimate the parameters of a primary generalized linear mixed model (GLMM) that accounts for baseline and/or time-dependent covariates. The misclassification process is modeled via a second generalized linear model that captures variations in sensitivity and specificity parameters according to time and a set of subject-specific covariates that may or may not overlap with those in the primary model. Simulation studies demonstrate the precision and validity of the proposed method. An application is presented based on longitudinal assessments of bacterial vaginosis conducted in the HIV Epidemiology Research (HER) Study.

Keywords

Bias; Differential misclassification; Nonlinear mixed model; Validation

1. Introduction

Many researchers have investigated the impact of binary variable misclassification on statistical inference. It is widely known that misclassification can lead to severe bias as well

*li.tang@stjude.org.

as reduced statistical efficiency (Barron (1977), Copeland *et al.* (1997), Neuhaus (1999) and Carroll *et al.* (2006)). There is broad literature on methods to correct for response misclassification, mainly in the context of ordinary logistic regression assuming known misclassification probabilities or the availability of validation data under certain assumptions (Green (1983), Greenland (1988), Marshall (1990), Brenner and Gefeller (1993), Magder and Hughes (1997), Morrissey and Spiegelman (1999), Carroll *et al.* (2006) and Greenland (2008)). In a generalized linear model context, Neuhaus (1999) quantified the magnitude of the bias when the response is misclassified and showed that the class of generalized linear models shares a closure property when misclassification probabilities are independent of covariates. That is, the error-prone responses continue to follow a generalized linear model, however with a modified link.

Recent literature illustrates approaches to incorporate validation data into the estimation of regression coefficients when the outcome is differentially misclassified, via the use of a Bayesian framework (Paulino *et al.* (2003), McInturff *et al.* (2004) and Gerlach and Stamey (2007)), nonparametric kernel methods (Pepe (1992)), or likelihood-based methods (Carroll *et al.* (2006)). Holcroft *et al.* (1997) proposed a 3-stage validation design, using inverse probability weighting. Given efficient optimization tools available in commercial software, Lyles *et al.* (2011) demonstrated a highly accessible ML implementation to correct for differentially misclassified binary outcomes in ordinary logistic regression based on internal validation data.

Despite a wide range of choices for potential correction methods, most attention has focused on the case where there is no repeated measurement of the response. However, longitudinal studies are common in practice. For example, in the observational HIV Epidemiology Research (HER) study, semiannual diagnoses of bacterial vaginosis (BV) were made on HIV infected or uninfected but at risk women (Smith *et al.* (1997)). One scientific question of interest is to identify important risk factors for BV in this subpopulation. The major complication of the analysis lies in error-prone diagnoses of BV (Lyles *et al.* (2011)). Thus, an efficient and computationally accessible method to adjust for differential misclassification in correlated binary outcomes is in demand, especially with a view toward large-scale epidemiological studies.

Neuhaus (2002) proposed a framework for implementing population-averaged (GEE) and cluster-specific generalized linear mixed model (GLMM) analyses when misclassification probabilities are either known or unknown, but fixed and independent of covariates. Taking advantage of a closure property (Neuhaus (1999)), he noted that ML estimates can theoretically be obtained when misclassification probabilities depend on covariates; however, the practical implication is that identifiability issues may arise without further assumptions in the spirit of sensitivity analysis. Subsequently, Lyles *et al.* (2005) examined the case of matched-pair 2×2 tables in a longitudinal study. When pairwise correlated responses are measured with error, they extended the idea of McNemar's test by incorporating external or internal validation data to estimate the paired-data odds ratio.

In the current work, we focus on longitudinal studies with repeatedly measured error-prone responses. Focusing on the GLMM framework, we integrate and extend the previous work

of Neuhaus (2002) and Lyles *et al.* (2005) to adjust for potentially complex differential misclassification mechanisms impacting a correlated binary outcome. The key to this added flexibility in the adjustment process is the availability of an internal validation subsample at different study time points, as is the case in a motivating application involving longitudinal binary assessments of BV in the HER study. This application clearly demonstrates how one could reach misleading conclusions without correcting for misclassification, and stresses the validity gains through the use of the proposed approach. Our objective is to make efficient ML methods accessible in this context, by taking advantage of optimization tools available in modern statistical software. The application to the HER study and simulation studies provide evidence of the serious biases that can be incurred without adequately modeling differential misclassification, while also demonstrating the performance of the proposed internal validation data-based approach.

2. Methods

2.1 Notation

Let Y_{ij} be the true response of interest for subject i at the j th occasion (e.g., study visit; $j=1, \dots, J_i$), with $Y_{ij}=1$ if disease is present and $Y_{ij}=0$ otherwise. The response depends on a set of covariates $\mathbf{X}_{ij}=(X_{ij1}, \dots, X_{ijp})^T$. We assume the following generalized linear mixed model:

$$g\{Pr(Y_{ij}=1|\mathbf{X}_{ij}, \mathbf{U}_i)\} = \beta_0 + \mathbf{X}_{ij}^T \boldsymbol{\beta} + \mathbf{Z}_{ij}^T \mathbf{U}_i, \quad (1)$$

in which g is an arbitrary link, $\boldsymbol{\beta}=(\beta_1, \dots, \beta_p)^T$ is a parameter vector, $\mathbf{Z}_{ij}=(Z_{ij1}, \dots, Z_{ijq})^T$ is a regressor vector for random effects, and $\mathbf{U}_i=(u_{i1}, \dots, u_{iq})^T$ is a subject-specific random effect vector generated from the (possibly multivariate) distribution $f(\mathbf{U})$. Conditional on $\mathbf{U}_i$, we assume that responses within subject i are conditionally independent (Breslow and Clayton (1993)), so that the likelihood for Y_{ij} can be fully specified as a function of β_0 , $\boldsymbol{\beta}$, $\mathbf{U}_i$ and parameters involved in $f(\mathbf{U})$. Henceforth, we make the common GLMM assumption that $\mathbf{U}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.

When Y_{ij} is potentially misclassified, we assume that information on error-prone responses is collected instead. Y_{ij}^* relates to Y_{ij} via the bridge of familiar diagnostic properties known as sensitivity ($SE=Pr(Y^*=1|Y=1)$) and specificity ($SP=Pr(Y^*=0|Y=0)$) in epidemiology. To model misclassification probabilities as a function of covariates, we introduce a secondary generalized linear model with an arbitrary link g' that need not be the same as the link g in (1):

$$g'\{Pr(Y_{ij}^*=1|Y_{ij}, \mathbf{C}_{ij}, \mathbf{U}_i^*)\} = \gamma_0 + \mathbf{C}_{ij}^T \boldsymbol{\gamma} + \gamma_{r+1} \mathbf{Y}_{ij} + \mathbf{W}_{ij}^T \mathbf{U}_i^*. \quad (2)$$

In (2), $\mathbf{C}_{ij}=(C_{ij1}, \dots, C_{ijq})^T$ denotes covariates that affect the response misclassification process. Similarly as in (1), $\boldsymbol{\gamma}$ is an r -th dimensional misclassification parameter vector, $\mathbf{W}_{ij}$ is a regressor vector for random effects, and $\mathbf{U}_i^*$ is a subject-specific random effect vector.

$$[U_i^T, U_i^{*T}]^T \sim N \left(\mathbf{0}, \begin{bmatrix} \Sigma \\ \Psi \\ \Psi \\ \Sigma^* \end{bmatrix} \right),$$

We further assume a joint distribution as U_i^* and U_i are independent. The variables in C_{ij} may or may not be a subset of those in X_{ij} , and the subscripts in C_{ij} indicate that misclassification rates depend on potentially time-varying subject-specific information. In particular, we define

$$SE(C_{ij}) = Pr(Y_{ij}^* = 1 | Y_{ij} = 1, C_{ij}, U_i^*) \quad \text{and} \quad SP(C_{ij}) = Pr(Y_{ij}^* = 0 | Y_{ij} = 0, C_{ij}, U_i^*).$$

These assumptions imply that the misclassification process may be correlated within the same subject, and may also be correlated with the response process.

Model (2) represents a highly flexible process, hereafter denoted as “general misclassification”. The setup of (2) allows AIC-based (Akaike (1974)) model evaluations to assess whether $\Psi = \mathbf{0}$ and whether U_i^* is necessary, and further to select variables that impact these key diagnostic properties (i.e., addressing the null hypothesis $H_0: \gamma_1 = \dots = \gamma_r = 0$). Under simple scenarios, for example, if only one parameter is involved in the covariance, a likelihood ratio test (LRT) can also be applied to test whether $\Psi = \mathbf{0}$ by comparing the test statistic to the distribution χ_1^2 . Model selection should be performed to ensure the inclusion of important covariates C in (2), as failing to do so may result in invalid estimates of regression coefficients in the main model (1).

An initial special case of (2) incorporates the assumption that the misclassification and response processes are uncorrelated (i.e., $\Psi = \mathbf{0}$), denoted as “independent correlated differential misclassification”. A second and more specific special case occurs when the misclassification process involves no correlation (i.e., $\Sigma^* = \mathbf{0}$), while SE and SP are differential by C . This is hereafter referred to as “uncorrelated differential misclassification”, and mimics the typical differential process discussed in most prior literature. In the latter case, (2) reduces to

$$g' \left\{ Pr(Y_{ij}^* = 1 | Y_{ij}, C_{ij}) \right\} = \gamma_0 + C_{ij}^T \gamma + \gamma_{r+1} Y_{ij}. \quad (3)$$

With (3) we assume that SE and SP for subject i at time point j depend only on his/her true response and covariate information at that time point, and that the misclassification processes for different occasions within the same subject are conditionally independent given that information. Specifically, we assume

$$Pr(Y_{ij}^* = y_{ij}^* | Y_i, C_i) = Pr(Y_{ij}^* = y_{ij}^* | Y_{ij}, C_{ij}).$$

In many practical situations, such conditional independence is intuitively defensible.

A more simplified special case of (3) is the following:

$$SE = Pr(Y_{ij}^* = 1 | Y_{ij} = 1) \quad \text{and} \quad SP = Pr(Y_{ij}^* = 0 | Y_{ij} = 0), \quad (4)$$

where both SE and SP are fixed constants across all subjects and occasions. In other words, $Pr(Y_{ij}^* = y_{ij}^* | \mathbf{Y}_i, \mathbf{C}_i) = Pr(Y_{ij}^* = y_{ij}^* | Y_{ij} = y_{ij})$, where $\mathbf{C}_i$ represents any set of time-dependent and time-independent subject-specific characteristics. This case is referred to widely, and in what follows, as “nondifferential misclassification”.

In the following sections, we focus on the case where $SE(C_{ij})$ and $SP(C_{ij})$ are estimable via validation data. Otherwise, one could conduct sensitivity analysis by supplying a range of possible values for these parameters while assuming “nondifferential misclassification” or “uncorrelated differential misclassification”.

2.2 Validation Sampling Schemes

In external validation sampling, data are collected independently of a current study sample (i.e., derived from previous similar studies or literature). To incorporate such information, one must assume “transportability”, i.e., that the misclassification probabilities operating in the external validation sample are the same as those operating in the current study (Carroll *et al.* (2006)). This assumption is questionable and often unverifiable in practice. In addition, only information on (Y_i, Y_i^*) would typically be available from an external validation study; note the lack of any ‘j’ subscript corresponding to time points in the main study. Thus, in most cases, the use of external validation data forces a fully nondifferential misclassification assumption because other information is not available.

In contrast, internal validation involves using a gold standard diagnostic technique to measure the true response for a randomly selected subsample of study participants. Benefits of such a design include avoidance of the transportability assumption, improved efficiency, and the flexibility of allowing for differential misclassification. In what follows, we assume the presence of supplemental data based on longitudinally-implemented internal validation subsampling.

2.3 Main / Internal Validation Study-Based Analysis under General Misclassification

Let $i (=1, \dots, n_m)$ index the subjects in the ‘main’ study for whom only the error-prone responses Y_{ij}^* and the covariates $(\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{C}_{ij}, \mathbf{W}_{ij})$ are measured. For the i th subject at occasion j , the likelihood conditional on random effects is as follows:

$$Pr(Y_{ij}^* | \mathbf{X}_{ij}, \mathbf{C}_{ij}, \mathbf{Z}_{ij}, \mathbf{W}_{ij}, U_i, U_i^*) = \sum_{Y_{ij}=0}^1 Pr(Y_{ij}^* | Y_{ij}, \mathbf{C}_{ij}, \mathbf{W}_{ij}, U_i^*) Pr(Y_{ij} | \mathbf{X}_{ij}, \mathbf{Z}_{ij}, U_i)$$

Thus the main study likelihood can be derived as

$$L_m = \prod_{i=1}^{n_m} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \prod_{j=1}^{J_i} \left[\{ (1 - SP(C_{ij})) + (SE(C_{ij}) + SP(C_{ij}) - 1) \times Pr(Y_{ij}=1 | \mathbf{X}_{ij}, \mathbf{Z}_{ij}, U_i) \}^{y_{ij}^*} \times \{ SP(C_{ij}) - (SE(C_{ij}) -$$

With regard to internal validation subsampling, one approach is to collect data on both the primary and the error-prone outcome [i.e., (Y_{ij}^*, Y_{ij})] for a random subsample of study subjects. That is, each validation subject contributes data on both the true and error-prone response at each time point, along with data on covariates (X_{ij}, Z_{ij}, C_{ij}) . While other validation sampling strategies are available (see Discussion), we proceed with the aforementioned scheme in what follows.

Let n_v be the number of subjects randomly chosen for validation. Without loss of generality, assume the data are sorted such that the first n_m subjects form the main study sample, and the following n_v subjects comprise the internal validation subset. The likelihood contribution from the validation sample is

$$L_v = \prod_{i=n_m+1}^{n_m+n_v} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \left\{ \prod_{j=1}^{J_i} Pr(Y_{ij}^* = y_{ij}^* | Y_{ij}, C_{ij}, W_{ij}, U_i) Pr(Y_{ij} = y_{ij} | X_{ij}, Z_{ij}, U_i) f(U, U^*) dU dU^* \right\}. \quad (6)$$

The full likelihood incorporating internal validation data is then proportional to $L_m \times L_v$.

2.4 Estimation

The likelihood for data collected under the proposed main/internal validation design can be computed by integrating out the random effects. Recommended numerical integration techniques include adaptive Gaussian quadrature (Pinheiro and Bates (1995)) and the first-order method (Beal and Sheiner (1988)). The full likelihood can then be maximized via quasi-Newton optimization, and standard errors obtained numerically by close approximation of the Hessian matrix. Such techniques are well developed in standard software packages. Our simulations and application to the HER study were carried out using the NLMIXED procedure in SAS 9.2 (SAS Institute, Inc. (2004)). Corresponding programs are readily adaptable, and are available from the authors. The computational time of the proposed approach depends on the complexity of the model (refer to Section 3 for further details). All analyses and simulations were run on a Dell Precision T5500 with Intel® Xeon® CPU X5687@3.60GHz (2 processors), a 4.00GB RAM and a 64-bit operating system.

3. Applications to HER Study

3.1 BV Data in HER Study

Our motivating applications are taken from the HIV Epidemiology Research (HER) study, a multi-center prospective cohort study with a total of 1310 women enrolled in four U.S. cities from 1993 to 1995 (Smith *et al.* (1997)). Among them, 871 HIV-infected and 439 uninfected but at risk women received a series of semi-annual diagnoses and assessments.

The question of interest is to assess the prevalence of BV when adjusting for necessary covariates. BV was measured by a clinically-based (CLIN) and a laboratory-based (LAB) method. CLIN diagnosis of BV was based on a modified set of Amsel criteria (Amsel *et al.* (1983)), which is error-prone but accessible. The more expensive and labor-intensive method, LAB, makes diagnoses via a sophisticated Gram-staining technique (Nugent *et al.*

(1991)), and serves as an arguable gold standard. An important feature of the HER study data is that both CLIN and LAB diagnoses were recorded at each visit. This facilitates a unique example, in that we can directly compare corrected results utilizing a validation subsample head-to-head versus both uncorrected and gold-standard-based results. We are thus able to make observations about both the validity and the efficiency of the validation-based approach in the context of the HER study, while informing future longitudinal studies about the benefits of validating a subsample under outcome misclassification. Previous analyses documented potentially complex differential misclassification in the HER study (Lyles *et al.* (2005) and (2010)), allowing us to reliably compare the performance of various misclassification models.

When initially applying the proposed methodology to estimate a paired data OR (without covariates), we obtained an identical result to that given in Lyles *et al.* (2005) via taking the same proportions of main and internal validation data at the 1st and 4th semi-annual HER study visits (results not shown). Henceforth, we focus upon the extension to the covariate-adjustment setting.

3.2 Application 1: Pairwise Covariate-adjusted Case

In the first application, we use data from black, white and Hispanic patients from the 1st and the 4th semi-annual HER study visits. We include 870 patients aged greater than or equal to 25 years old at enrollment, who had CLIN and LAB BV as well as relevant BV risk factors measured at both visits. Potential covariates associated with BV status include race, HIV status (negative or positive), HIV risk group (via sexual contact or via intravenous drug use (IDU)), and age in years. The median age at enrollment was 36.0 years. The study sample consists of 530 blacks (60.9%), 207 Caucasians (23.8%) and 133 Hispanics (15.3%). For model fitting purposes, Hispanics were combined with Caucasians after initial analysis suggesting similar BV prevalence for these two groups. 587 women were HIV positive (67.5%), and 465 were in the IDU group (53.5%). 227 women were from the New York site (26.1%), 196 from the Michigan site (22.5%), 225 from the Maryland site (25.9%), and 222 from the Rhode Island Site (25.5%). At the 1st semi-annual HER study visit, the estimated BV prevalence was 33.8% by the CLIN method and 46.3% by the LAB method. At the 4th visit, CLIN produced a crude BV prevalence estimate of 25.2%, in contrast to the LAB estimate of 40.9%. Thus, the CLIN method appears to significantly underestimate the prevalence of BV in the sample.

Crude SE and SP estimates in the sexual contact HIV risk group were 0.46 and 0.94 respectively, after combining data from the two visits. However, in the IDU risk group, SE and SP were estimated as 0.63 and 0.88. Thus, misclassification appears highly differential with respect to HIV risk group (Riskgrp). SE and SP also change over time, with crude (SE, SP) estimates at visit 1 of (0.59, 0.88) as opposed to (0.52, 0.94) at visit 4. Similarly, crude estimates combining data across visits suggest higher SE and lower SP in the HIV negative and at Maryland (MD) and Rhode Island (RI), as opposed to positive and at New York (NY) and Michigan (MI). All of these descriptive statistics suggest a complex differential misclassification process that could not be unraveled without internal validation data collected across visits.

Let Y_{ij} denote BV status for subject i at the j th time point, where $j=1, 4$. After preliminary model selection using the LAB data, we assume the following primary logistic model:

$$\text{logit} \{ \Pr(Y_{ij}=1 | \mathbf{X}_{ij}, u_i) \} = \beta_0 + \beta_1 \text{HIV pos} + \beta_2 \text{Age} + \beta_3 \text{Riskgrp} + \beta_4 \text{Race} + \beta_5 \text{Vst1} + u_i. \quad (7)$$

Visit 4 is considered the reference visit. A random subsample of 200 patients were selected for validation as described in Section 2.3. We then fit six models. The first is the ideal model with BV status measured by the gold-standard (LAB) method (Y_{ij}) as the response. The second is the “naïve” model with BV status measured by the error-prone CLIN method as the response, i.e., replacing Y_{ij} by Y_{ij}^* in (7).

For the joint model to adjust for misclassification via ML, we fit (7) and (8) simultaneously:

$$\text{logit} \left\{ \Pr \left(Y_{ij}^*=1 | Y_{ij}, \mathbf{C}_{ij}, u_i^* \right) \right\} = \gamma_0 + \gamma_1 \text{HIV pos} + \gamma_2 \text{Riskgrp} + \gamma_3 \text{Race} + \gamma_4 \text{Vst1} + \gamma_5 \text{NY} + \gamma_6 \text{MI} + \gamma_7 \text{MD} + \gamma_8 Y_{ij} + u_i^*. \quad (8)$$

RI serves as the reference site. The random effects u_i^* accommodate within-subject correlations in misclassification across the two semi-annual visits. It is assumed that

$$\begin{bmatrix} u_i & u_i^* \end{bmatrix}' \sim N \left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 & \psi \\ \psi & \sigma_2^2 \end{bmatrix} \right). \quad (8)$$

Model (8) was selected based on validated observations, representing general misclassification. To produce analyses assuming independent correlated differential misclassification, uncorrelated differential misclassification, and nondifferential misclassification, we restrained $\psi=0$, excluded u_i^* from (8) and removed all covariates except y_{ij} from (8), step by step. Although formal selection of the secondary model is not our primary focus, (8) is supported by univariate preliminary investigations of the misclassification process.

Table 1 reveals that the naïve error-prone model and the gold-standard model differ markedly in the magnitude of the estimated OR for HIV risk group (1.62 for the ideal analysis, 2.73 for the naïve) and in the directionality of the HIV status association (1.02 and non-significant for the ideal analysis, 0.70 and significant for the naïve) This highlights the need to adjust for outcome misclassification in studies like the HER study that utilize CLIN BV assessments.

Table 2 summarizes results based on the main/internal validation data likelihood, first allowing for general misclassification (Model 1), then assuming independent correlated differential misclassification (Model 2), uncorrelated differential misclassification (Model 3) and nondifferential misclassification (Model 4). The independent correlated differential misclassification model (Model 2) has the smallest AIC, and was chosen as our correction model. This agrees with the LRT result (statistic for testing the need to include the covariance=0.7, $p=0.40$). Note that when accounting for differential misclassification with or without correlation in misclassification within the same subject as in Models 1, 2 and 3, interpretations are very similar to those obtained by fitting the gold-standard model with regard to the magnitudes and directionality of the estimated ORs. In contrast, when inappropriately assuming nondifferentiability (Model 4), the point estimates differ markedly

from the gold-standard analysis (for example, the estimated OR for HIV status is in the same direction as for the naïve analysis). This highlights the importance of modeling SE/SP differentially when nondifferentiality could be suspect. Interestingly, results from Models 1-3 are very similar in terms of the magnitudes of point estimates, which is consistent with our application 2 in Section 3.3 and with empirical results in Section 4. However, one may argue on the basis of AIC (see table footnotes) that Models 1 and 2 fit slightly better than Model 3.

Based on the joint ML analysis summarized in Table 2, using the independent correlated differential misclassification model, SE and SP are found associated with HIV status, risk group, time index, race and sites. SE tends to be higher and SP lower in HIV negatives

$(\hat{Y}_1 = -0.38, p=0.04)$, subjects at risk via IDU $(\hat{Y}_2 = 0.38, p=0.04)$, black women $(\hat{Y}_3 = 0.51, p=0.02)$, at visit 1 $(\hat{Y}_4 = 0.48, p=0.005)$, and at the MD and RI sites $(\hat{Y}_5 = -0.64, p=0.003; \hat{Y}_6 = -1.51, p<0.001; \hat{Y}_7 = 0.43, p=0.07)$. These findings are consistent with our crude estimates.

3.3. Application 2: Longitudinal Analysis with More Than 2 Time Points

In the second application, we consider 706 patients older than 25 at the time of enrollment who had complete data on BV status from the 1st through the 4th semi-annual visits. The median age at enrollment remained at 36.0 years. There were 425 blacks (60.2%) in the sample. 487 women were HIV positive (69.0%), and 392 were in the IDU group (55.5%). 180 were from NY (25.5%), 135 from MI (19.1%), 197 from MD (27.9%), and 194 from RI (27.5%). Crude BV prevalence estimates were relatively stable over time based on LAB (Y) measurements, but decreased across visits based on CLIN (Y^*) assessments. At each visit, prevalence was markedly higher based on LAB as opposed to CLIN.

We consider three types of models: the ideal analysis with the gold-standard (LAB) as the response, the naïve model with BV status measured by the error-prone CLIN method as the response, and joint models fit using the proposed ML approach. On the basis of preliminary model selection, for the “gold-standard” model, we fit

$$\text{logit} \{Pr(Y_{ij}=1|\mathbf{X}_{ij}, u_i)\} = \beta_0 + \beta_1 HIV_{pos} + \beta_2 Age + \beta_3 Risk_{grp} + \beta_4 Race + \beta_5 V_{st1} + \beta_6 V_{st2} + \beta_7 V_{st3} + u_i. \quad (9)$$

Again, we consider visit 4 as the reference visit. To adjust for differential misclassification based on the main/internal validation design, we fit (9) jointly with the following SE/SP model:

$$\begin{aligned}
& \text{logit} \left\{ \Pr \left(Y_{ij}^* = 1 | Y_{ij}, C_{ij}, u_i^* \right) \right\} \\
&= \gamma_0 + \gamma_1 \text{HIV} \text{pos} \\
&+ \gamma_2 \text{Riskgrp} \\
&+ \gamma_3 \text{Race} \\
&+ \gamma_4 \text{Vst1} \\
&+ \gamma_5 \text{Vst2} \\
&+ \gamma_6 \text{Vst3} \\
&+ \gamma_7 \text{NY} \\
&+ \gamma_8 \text{MI} \\
&+ \gamma_9 \text{MD} \\
&+ \gamma_{10} Y_{ij} + u_i^*.
\end{aligned} \tag{10}$$

In (10), we again treat RI as the reference site. Allowing for general misclassification

(Model 1), we assume u_i and u_i^* are jointly distributed as $N \left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 & \psi \\ \psi & \sigma_2^2 \end{bmatrix} \right)$. Assuming independent correlated differential misclassification (Model 2), the joint distribution of u_i

and u_i^* follows $N \left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$. We may further assume uncorrelated differential misclassification (Model 3), which implies that u_i^* is removed from (10). When assuming non-differential misclassification (Model 4), we removed u_i^* and all covariates except y_{ij} in (10). We randomly selected 200 patients into the internal validation subsample. The remaining 506 patients comprised the main study sample, and contributed only Y^* (not Y) observations.

Table 3 summarizes the fit of the ‘naïve’ and ‘ideal’ models. Overall, results are similar to those in Table 1, in which only data from visits 1 and 4 were analyzed. The naïve model produces an estimated OR for HIV status in the opposite direction as in the gold-standard model (estimated OR=0.75 vs. 1.04), and an inflated OR estimate for HIV risk group (3.08 vs. 1.98). There are also noteworthy differences in the estimated visit effects.

Table 4 provides the results of ML analyses based on the main/internal validation design. Allowing for general misclassification (Model 1) or independent correlated differential misclassification (Model 2) via (10), the results are similar to those from the ideal analysis except for predictably larger standard errors. The independent correlated differential misclassification model was selected via the AIC criterion (see footnotes), which was also supported by the LRT (statistic for testing the need to include the covariance=1.4, $p=0.24$). SE and SP were significantly associated with race, HIV status, HIV risk group, time index and site index based on (10) in the joint ML analyses. Again SE tends to be higher in blacks, patients at risk via IDU, HIV negatives, at visits 1 and 2, and at RI and MD. When assuming uncorrelated differential misclassification (Model 3), estimates are similar to those from

more complicated models. When erroneously assuming nondifferentiability (Model 4), however, the estimated coefficients closely resemble those based on the naïve analysis.

4. Simulation Studies

In this section, we describe simulation studies to assess the performance of the proposed approaches. We confine attention to the case where g and g' in (1) and (4) are logit links; however, other links can readily be used.

In Table 5, we examine the performance of the MLEs based on the proposed approach in a hypothetical longitudinal study with four time points, with general response misclassification. The main model (11) includes a binary covariate X_1 generated as Bernoulli (0.5), and a continuous covariate X_2 generated as $N(0,4)$. True values of $(\beta_0, \beta_1, \beta_2)$ were (0, 1, 0.5).

$$\text{logit} \{ \Pr(Y_{ij}=1 | \mathbf{X}_{ij}, u_i) \} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i. \quad (11)$$

A secondary logistic model (12) was used to allow misclassification probabilities to depend on individuals' covariate information, as follows:

$$\text{logit} \left\{ \Pr(Y_{ij}^*=1 | Y_{ij}, \mathbf{T}, \mathbf{X}_{ij}, u_i^*) \right\} = \gamma_0 + \gamma_1 I(T_{ij}=2) + \gamma_2 I(T_{ij}=3) + \gamma_3 I(T_{ij}=4) + \gamma_4 X_{i1} + \gamma_5 Y_{ij} + u_i^*, \quad (12)$$

where I is an indicator function with a value of 1 if the condition described in parentheses is satisfied. The vectors (u_i, u_i^*) were simulated from a bivariate normal distribution with a

covariance matrix of $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. True values of $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$ were taken to be $(-3, 0.05, 0.2, 0.4, 1.5, 3)$. 500 simulations were conducted, each generating 1000 “subjects”. In each sample, 800 of these provided only main study observations, and 200 were randomly assigned for internal validation (see Section 2.3).

The summary in Table 5 suggests that fitting the independent correlated differential misclassification version of (12) that ignores $\text{cov}(u_i, u_i^*)$ does not lead to apparent bias in estimating primary parameters of interest, although the ignored covariance produces a perception of additional variability in u_i and u_i^* , with inflated $\hat{\sigma}_{u_i}^2$ and $\hat{\sigma}_{u_i^*}^2$. In contrast, if one assumes uncorrelated differential misclassification or completely nondifferential misclassification, empirical evidence suggests that the validity of $\hat{\beta}_2$ still holds, while $\hat{\beta}_1$ may no longer be valid. The explanation for this observation lies in the fact that X_1 is also part of the misclassification model (12), so that mis-specifying this model may lead to bias in estimating β_1 .

In practice, when there is sufficient validation information, we suggest one starts with the general or independent correlated differential misclassification model to help ensure the validity of the corrected results. Unless suggested by the data, one should not rely heavily on oversimplified models, especially the one assuming completely nondifferential

misclassification. Such oversimplification could lead to severe biases that are comparable to those incurred when performing naïve analyses. The “Selected Model” in Table 5 refers to the one of the four main/validation study misclassification models that was selected for each simulation run on the basis of the lowest AIC (see Section 2.1). The AIC criterion selected the correct (general misclassification) model 93% of the time for the simulations summarized in Table 5. Note that this strategy preserves overall validity and satisfactory 95% CI coverage, and can readily be applied in practice to guide model selection in this context.

In Table 6, we also evaluated a setting in which data were generated under the same conditions as in Table 5 except $cov(u_i, u_i^*) = 0$. In this case, AIC reliably selected the correct (independent correlated differential) misclassification model 87% of the time, yielding results similar to those from the ideal gold-standard model. Interestingly, AIC selected the general model 13% of the time and never selected an oversimplified model, indicating that this straightforward model selection strategy performs well.

In Table 7, we assessed the performance of the proposed approach when data were generated under the uncorrelated differential misclassification model by removing u_i^* from (12). Other parameters stayed the same. The correct model was selected 91.8% of the time, while more general models were chosen 8.2% of the time. Again, the oversimplified nondifferential model that led to severe biased estimates was never selected. The correction model yielded similar results as those observed in tables 5 and 6.

Tables 5, 6 and 7 suggest little in the way of efficiency benefits when estimating the primary parameters under a simpler misclassification model than the one used to generate the data. Therefore, it may be safest to fit the general or the independent correlated differential misclassification model, if the validation size is adequate, to ensure the validity of the correction. Meanwhile, in order to facilitate the optimization, it may be helpful to apply estimates from simplified models as initial values.

5. Discussion

Our work differs in several ways from most previous treatments of outcome misclassification in longitudinal studies. First, we explicitly provide the form of a joint main/internal validation study likelihood for repeatedly measured misclassified responses. Second, we do not restrict misclassification probabilities to be non-differential, instead emphasizing the benefits of directly modeling covariate effects in the misclassification process. Third, we provide an accessible computational approach to optimize the likelihood using standard software, with sharable programs available. Finally, we provide a means by which the usual assumption of conditional independence in the misclassification process over time or across cluster members can be relaxed. We strongly recommend the internal validation strategy when possible, to avoid the unverifiable “transportability” assumption, gain precision in estimation, and, especially, allow for more flexibility in modeling misclassification probabilities. The proposed approach also makes it possible for one to check whether it is necessary to adjust for differential misclassification via an AIC-based model selection approach. Nevertheless, for validity purposes, it may be safer to employ a

more general SE/SP model. Both simulation studies and applications to the HER study indicate that internal validation data are crucial for unraveling the potentially deleterious effects of differential misclassification.

Throughout, we have assumed random effects are normally distributed. It is known (e.g., Litière *et al.* (2008)) that misspecifying the distribution of random effects could potentially lead to biased estimates. The approach presented may be extended to accommodate non-normal random effects without conceptual difficulty; however, relaxing the normality assumption requires more complex numerical optimization techniques and could invite alternative Bayesian solutions, which could provide motivation for future research. Attention should also be paid to the choice of the link in both the response and misclassification models. For our modeling purposes, the popular logit link offered a natural choice for both. However, prior work has emphasized that misspecification of a link function can lead to biased estimates in the generalized linear model setting (Pregibon (1980) and Koenker and Yoon (2009)). The proposed approach can be extended to other links in principle. In this regard, we re-fit the selected models in our HER study applications using probit and complementary log-log links. Convergence was obtained, and the resulting interpretations in terms of statistically significant covariates were very similar to those presented based on logit links in Tables 2 and 4 (results not shown).

Theoretically, multiple random effects and nested levels (i.e., random site effects in a multicenter study) can be introduced into the proposed approach; however, including more random effects can also result in computational challenges. Our user-friendly program is implemented via SAS NLMIXED, which does not support multilevel models (SAS Institute Inc. (2004)). We have, however, included sites as fixed effects in our real data applications, which is a defensible approach in the HER study.

In our applications, we confine our attention to the situation where model selection is based only on statistical considerations, such as AIC. However, in practice, we encourage users to incorporate prior scientific knowledge regarding the misclassification process into model selection, to ensure the selected model is scientifically sound. In addition, further consideration of validation study designs, including validation sample size considerations, could be of great use and general interest. In this direction, Tang (2012) discusses alternatives to the validation study design applied here. Also along these lines, a sample simulation program investigating required subsample size for maintaining analytic validity with pre-specified primary and misclassification parameters is available from the author upon request. One may also be interested in a formal consideration of cost-efficiency via the validation sampling strategy, by extending prior work (Spiegelman and Gray (1991) and Lyles *et al.* (2005)) to the longitudinal setting addressed here. Furthermore, our approach relies on a gold-standard method in order to perform the correction for misclassification. In practice, there may not be a gold-standard technique available. In that case, replicates or multiple imperfect diagnostic tests might be used to develop a valid correction method, although some of the critical flexibility for modeling differential misclassification via internal validation data will likely be lost. Future work could also include developing alternative semi-parametric methods such as generalized additive models, with emphasis on

accommodating (via study design and analysis) potentially complex differential misclassification patterns.

ACKNOWLEDGEMENTS

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 HER Study was supported by the Centers for Disease Control and Prevention (CDC): U64/CCU106795, U64/CCU206798, U64/CCU306802, U64/CCU506831. The content is solely the responsibility of the authors and does not necessarily represent the official position of the National Institutes of Health or the CDC. The authors especially thank the participants of the HER Study and the Study Research Group. The HER Study Research Group 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, Ph.D., 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 Davenney, Ph.D. from the National Institute of Drug Abuse.

REFERENCES

- Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974; 19:716–723.
- 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]
- Barron BA. The effects of misclassification on the estimation of relative risk. *Biometrics*. 1977; 33:414–417. [PubMed: 884199]
- Beal SL, Sheiner LB. Heteroskedastic Nonlinear Regression. *Technometrics*. 1988; 30:327–338.
- Brenner H, Gefeller O. Use of positive predictive value to correct for disease misclassification in epidemiologic studies. *American Journal of Epidemiology*. 1993; 138:1007–1015. [PubMed: 8256775]
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*. 1993; 88:9–25.
- Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Measurement Error in Nonlinear Models*. Second Edition. Chapman and Hall; London: 2006.
- Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology*. 1997; 105:488–495. [PubMed: 871121]
- Gerlach R, Stamey J. Bayesian model selection for logistic regression with misclassified outcomes. *Statistical Modelling*. 2007; 7:255–273.
- 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]
- Greenland S. Variance estimation of epidemiologic effect estimates under misclassification. *Statistics in Medicine*. 1988; 7:745–757. [PubMed: 3043623]
- Greenland S. Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. *Journal of Statistical Planning and Inference*. 2008; 138:528–538.
- Holcroft CA, Rotnitzky A, Robins JM. Efficient estimation of regression parameters from multistage studies with validation of outcomes and covariates. *Journal of Statistical Planning and Inference*. 1997; 65:349–374.
- Koenker R, Yoon J. Parametric links for binary choice models: a Fisherian-bayesian colloquy. *Journal of Econometrics*. 2009; 152:120–130.

- Litiere S, Alonso A, Molenberghs G. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine*. 2008; 27:3125–3144. [PubMed: 18069726]
- Lyles RH, Williamson JM, Lin HM, Heiling CM. Extending McNemar's test: estimation and inference when paired binary outcome data are misclassified. *Biometrics*. 2005; 61:287–294. [PubMed: 15737105]
- 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; 22:589–97. [PubMed: 21487295]
- Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*. 1997; 146:195–203. [PubMed: 9230782]
- Marshall RJ. Validation study methods for estimating proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*. 1990; 43:941–947. [PubMed: 2213082]
- McInturff P, Johnson WO, Cowling D, Gardner IA. Modeling risk when binary outcomes are subject to error. *Statistics in Medicine*. 2004; 23:1095–1109. [PubMed: 15057880]
- Morrissey MJ, Spiegelman D. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics*. 1999; 55:338–344. [PubMed: 11318185]
- Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*. 1999; 86:843–55.
- Neuhaus JM. Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*. 2002; 58:675–73. [PubMed: 12230004]
- 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]
- Paulino CD, Soares P, Neuhaus J. Binomial regression with misclassification. *Biometrics*. 2003; 59:670–675. [PubMed: 14601768]
- Pepe MS. Inference using surrogate outcome data and a validation sample. *Biometrika*. 1992; 79:355–365.
- Pinheiro JC, Bates DM. Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model. *Journal of Computational and Graphical Statistics*. 1995; 4:12–35.
- Pregibon D. Goodness of link tests for generalized linear models. *Applied Statistics*. 1980; 29:15–24.
- SAS Institute, Inc.. SAS/STAT 9.1 User's Guide. SAS Institute, Inc.; Cary, NC: 2004.
- 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]
- Tang, L. Analysis of Data with Complex Misclassification in Response or Predictor Variables by Incorporating Validation Subsampling. Department of Biostatistics and Bioinformatics, Emory University; 2012. (unpublished PhD dissertation)

Table 1

Change in BV prevalence for women between HER study visits 1 and 4 with covariate adjustment (Ideal and Naïve Analysis)

Analysis Based on Ideal Model Using Gold Standard LAB Method [†]			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	0.02(0.15)	1.02 (0.77, 1.36)	0.89
Age	-0.05(0.01)	0.95 (0.93, 0.97)	<0.001
Risk Group (IDU vs Sex)	0.48(0.14)	1.62 (1.23, 2.13)	0.001
Race (Black vs Other)	0.93(0.15)	2.53 (1.90, 3.37)	<0.001
Visit (Visit 1 vs. Visit 4)	0.29 (0.11)	1.34 (1.07, 1.66)	0.01
Analysis Based on Naïve Model Using Error-Prone CLTN Method ^{††}			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	-0.35(0.16)	0.70 (0.52, 0.96)	0.03
Age	-0.06(0.01)	0.94 (0.92, 0.96)	<0.001
Risk Group (IDU vs Sex)	1.01(0.16)	2.73 (2.01, 3.72)	<0.001
Race (Black vs Other)	1.08(0.16)	2.94 (2.14, 4.05)	<0.001
Visit (Visit 1 vs. Visit 4)	0.56 (0.12)	1.75 (1.37, 2.23)	<0.001

[†] ML based on (7).

^{††} ML based on (7) by replacing Y with Y^* .

Table 2

Change in BV prevalence for women between HER study visits 1 and 4 with covariate adjustment (Main/Internal Validation Analysis).

Main/Internal. General Misclassification[†]			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	0.02 (0.22)	1.02 (0.66, 1.59)	0.92
Age	-0.06 (0.02)	0.94 (0.91, 0.97)	<0.001
Risk Group (IDU vs Sex)	0.49 (0.22)	1.63 (1.07, 2.48)	0.02
Race (Black vs Other)	0.96 (0.23)	2.62 (1.67, 4.12)	<0.001
Visit (Visit 1 vs. Visit 4)	0.23 (0.20)	1.26 (0.85, 1.87)	0.24
Main/Internal, Independent Correlated Differential Misclassification^{††}			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	0.02 (0.23)	1.02 (0.65, 1.60)	0.93
Age	-0.06 (0.02)	0.94 (0.91, 0.97)	<0.001
Risk Group (IDU vs Sex)	0.50 (0.22)	1.65 (1.07, 2.54)	0.02
Race (Black vs Other)	0.96 (0.24)	2.61 (1.65, 4.15)	<0.001
Visit (Visit 1 vs. Visit 4)	0.24 (0.20)	1.27 (0.86, 1.87)	0.24
Main/Internal, Uncorrelated Differential Misclassification^{†††}			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	0.03 (0.25)	1.03 (0.64, 1.68)	0.89
Age	-0.06 (0.02)	0.94 (0.90, 0.97)	<0.001
Risk Group (IDU vs Sex)	0.48 (0.24)	1.62 (1.02, 2.58)	0.04
Race (Black vs Other)	1.08 (0.25)	2.94 (1.79, 4.84)	<0.001
Visit (Visit 1 vs. Visit 4)	0.26 (0.20)	1.29 (0.86, 1.93)	0.21
Main/Internal, Nondifferential Misclassification^{††††}			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	-0.30 (0.21)	0.74 (0.49, 1.12)	0.16
Age	-0.08 (0.02)	0.92 (0.89, 0.95)	<0.001
Risk Group (IDU vs Sex)	1.08 (0.21)	2.94 (1.94, 4.46)	<0.001
Race (Black vs Other)	1.40 (0.23)	4.06 (2.61, 6.32)	<0.001
Visit (Visit 1 vs. Visit 4)	0.57 (0.18)	1.77 (1.26, 2.49)	0.001

[†]SE and SP assumed to vary with the binary variables HIV risk group, HIV status, race, visit (Visit 4 as the reference level) and sites (RI as the reference level) via (8). $n_M=670$, $n_V=200$. $\hat{\sigma}_u^2=0.37$. $\hat{\sigma}_{u^*}^2=1.24$. $\hat{\psi}=0.26$. Computing time=40min. AIC=2334.0.

^{††}SE and SP assumed to vary via (8) assuming $cov(u_i, u_i^*)=0$. $\hat{\sigma}_u^2=0.41$. $\hat{\sigma}_{u^*}^2=1.51$. Computing time=32min. AIC=2332.7.

^{†††}SE and SP assumed to vary via (8) without the random effects term u_i^* . $\hat{\sigma}_u^2=0.94$. AIC=2341.9. Computing time=4 min.

^{††††}No covariates affecting SE and SP. $\hat{\sigma}_u^2=1.21$. Computing time=1min. AIC=2454.7.

Table 3

Change in BV prevalence for women from HER study visits 1 through visit 4 with covariate adjustment (Ideal and Naïve Analysis).

Analysis Based on Ideal Model Using Gold Standard LAB Method [†]			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	0.04(0.17)	1.04 (0.74, 1.46)	0.82
Age	-0.06(0.01)	0.94 (0.91, 0.96)	<0.001
Risk Group (IDU vs Sex)	0.68(0.16)	1.98 (1.44, 2.73)	<0.001
Race (Black vs Other)	1.20(0.17)	3.31 (2.37, 4.63)	<0.001
Visit (Visit 1 vs. Visit 4)	0.35 (0.13)	1.42 (1.09, 1.85)	0.01
Visit (Visit 2 vs. Visit 4)	0.15 (0.13)	1.17 (0.90, 1.52)	0.25
Visit (Visit 3 vs. Visit 4)	0.10 (0.13)	1.11 (0.85, 1.44)	0.46

Analysis Based on Naïve Model Using Error-Prone CLIN Method ^{††}			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	-0.28(0.15)	0.75 (0.56, 1.01)	0.06
Age	-0.07(0.01)	0.94 (0.91,0.96)	<0.001
Risk Group (IDU vs Sex)	1.13(0.15)	3.08 (2.30,4.12)	<0.001
Race (Black vs Other)	1.18(0.15)	3.25 (2.41,4.39)	<0.001
Visit (Visit 1 vs. Visit 4)	0.56 (0.14)	1.75 (1.33, 2.30)	<0.001
Visit (Visit 2 vs. Visit 4)	0.43 (0.14)	1.54 (1.17, 2.03)	0.002
Visit (Visit 3 vs. Visit 4)	0.25 (0.14)	1.28 (0.97, 1.69)	0.08

[†] ML based on (9).

^{††} ML based on (9) by replacing Y with Y^* .

Table 4

Change in BV prevalence for women from HER study visits 1 through visit 4 with covariate adjustment (Main/Internal Validation Analysis).

Main/Internal, General Misclassification[†]			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	0.33 (0.29)	1.39 (0.79, 2.43)	0.25
Age	−0.06 (0.02)	0.94 (0.90, 0.98)	0.001
Risk Group (IDU vs Sex)	0.64 (0.26)	1.90 (1.14, 3.17)	0.01
Race (Black vs Other)	0.80 (0.27)	2.23 (1.32, 3.76)	0.003
Visit (Visit 1 vs. Visit 4)	0.14 (0.23)	1.15 (0.73, 1.81)	0.55
Visit (Visit 2 vs. Visit 4)	0.07 (0.24)	1.07 (0.67, 1.69)	0.77
Visit (Visit 3 vs. Visit 4)	0.07 (0.23)	1.07 (0.68, 1.69)	0.75

Main/Internal, Independent Correlated Differential Misclassification^{††}			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	0.31 (0.29)	1.37 (0.77, 2.45)	0.28
Age	−0.06 (0.02)	0.94 (0.90, 0.97)	0.001
Risk Group (IDU vs Sex)	0.64 (0.27)	1.90 (1.12, 3.22)	0.02
Race (Black vs Other)	0.82 (0.27)	2.28 (1.33, 3.91)	0.003
Visit (Visit 1 vs. Visit 4)	0.15 (0.23)	1.16 (0.74, 1.82)	0.53
Visit (Visit 2 vs. Visit 4)	0.06 (0.24)	1.06 (0.67, 1.69)	0.80
Visit (Visit 3 vs. Visit 4)	0.07 (0.23)	1.07 (0.68, 1.69)	0.76

Main/Internal, Uncorrelated Differential Misclassification^{†††}			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	0.33 (0.28)	1.40 (0.80, 2.43)	0.24
Age	−0.06 (0.02)	0.94 (0.90, 0.97)	0.001
Risk Group (IDU vs Sex)	0.65 (0.26)	1.92 (1.15, 3.18)	0.01
Race (Black vs Other)	0.86 (0.26)	2.35 (1.40, 3.96)	0.001
Visit (Visit 1 vs. Visit 4)	0.14 (0.23)	1.15 (0.73, 1.81)	0.54
Visit (Visit 2 vs. Visit 4)	0.09 (0.24)	1.10 (0.57, 1.45)	0.69
Visit (Visit 3 vs. Visit 4)	0.08 (0.23)	1.08 (0.59, 1.46)	0.74

Main/Internal, nondifferential Misclassification^{††††}			
Variable	β(SE)	Estimated OR (95% CI)	P-value
HIV Status (Positive vs. Negative)	−0.33 (0.24)	0.72 (0.45, 1.16)	0.17
Age	−0.09 (0.02)	0.91 (0.88, 0.95)	<0.001
Risk Group (IDU vs Sex)	1.39 (0.24)	4.02 (2.53, 6.39)	<0.001
Race (Black vs Other)	1.57 (0.24)	4.79 (2.97, 7.73)	<0.001
Visit (Visit 1 vs. Visit 4)	0.52 (0.20)	1.68 (1.12, 2.50)	0.01
Visit (Visit 2 vs. Visit 4)	0.27 (0.20)	1.31 (0.88, 1.96)	0.18

Main/Internal, nondifferential Misclassification ^{††††}			
Variable	β (SE)	Estimated OR (95% CI)	P-value
Visit (Visit 3 vs. Visit 4)	0.12 (0.21)	1.13 (0.75, 1.69)	0.56

[†] SE and SP assumed to vary with the binary variables HIV risk group, HIV status, race, visit (Visit 4 as the reference level) and sites (RI site as the reference level) via (10). $n_m=507$, $n_v=200$. $\hat{\sigma}_u^2=3.13$. $\hat{\sigma}_{u^*}^2=0.46$. $\hat{\psi}=0.34$. Computing time=222min. AIC=3731.8.

^{††} SE and SP assumed to vary via (10) assuming $cov(u_i, u_i^*)=0$. $\hat{\sigma}_u^2=3.31$. $\hat{\sigma}_{u^*}^2=0.54$. Computing time=213min. AIC=3731.2.

^{†††} SE and SP assumed to vary via (10) without the random effects term u_i^* . $\hat{\sigma}_u^2=3.46$. AIC=3738.2. Computing time=15 min.

^{††††} No covariates affecting SE and SP. $\hat{\sigma}_u^2=4.10$. Computing time=2min. AIC=3912.8.

Table 5

Simulations comparing MLEs under internal validation design for logistic-mixed regression with general outcome misclassification and two covariates ($J=4$)[†]

Model	β_1		β_2		σ_u^2		$\hat{\sigma}_{u*}^2$		ψ	
	Mean (SD)	95% Coverage	Mean (SD)	95% Coverage	Mean (SD)	95% Coverage	Mean (SD)	95% Coverage	Mean (SD)	95% Coverage
Naive	0.90 (0.10)	82.8%	0.21 (0.03)	0	1.15 (0.15)					
Ideal	1.00 (0.11)	92.6%	0.50 (0.03)	96.6%	1.00 (0.14)					
General Misclassification ^{††}	0.99 (0.20)	93.6%	0.50 (0.05)	94.4%	1.00 (0.30)		1.01 (0.23)		0.52 (0.16)	
Independent Correlated Differential Misclassification ^{†††}	1.00 (0.20)	95.8%	0.49 (0.05)	93.6%	1.11 (0.31)		1.38 (0.26)			
Uncorrelated Differential Misclassification ^{††††}	1.07 (0.22)	90.2%	0.49 (0.05)	89.6%	2.41 (0.22)					
Nondifferential Misclassification ^{†††††}	1.53 (0.19)	16.4%	0.48 (0.05)	89.8%	2.39 (0.22)					
Selected Model	1.00 (0.20)	93.2%	0.50 (0.05)	93.2%	1.02 (0.36)		1.02 (0.23)		0.54 (0.16)	

[†] ML based on (11) and (12); 500 simulations under each set of conditions with $(\beta_0, \beta_1, \beta_2, \sigma_u^2, \sigma_{u*}^2, \psi) = (0, 1, 0.5, 1, 1, 0.5), (0, 1, 0.5, 1, 1, 0.5), (0, 1, 0.5, 1, 1, 0.5), (0, 1, 0.5, 1, 1, 0.5), (0, 1, 0.5, 1, 1, 0.5)$ and $X_2 \sim N(0.4), n_M=800, n_V=200$.

^{††} Converged 499 times.

^{†††} Converged 500 times.

^{††††} Converged 468 times.

^{†††††} Converged 476 times.

Table 6

Simulations comparing MLEs under internal validation design for logistic-mixed regression with general outcome misclassification and two covariates ($J=4$)[†]

Model	β_1		β_2		$\hat{\sigma}_u^2$	$\hat{\sigma}_{u*}^2$	ψ
	Mean (SD)	95% Coverage	Mean (SD)	95% Coverage	Mean (SD)	Mean (SD)	Mean (SD)
Naive	1.58 (0.10)	0	0.23 (0.02)	0	0.78 (0.12)		
Ideal	1.00 (0.10)	94.6%	0.50 (0.03)	95.8%	1.00 (0.14)		
General Misclassification ^{††}	1.00 (0.20)	92.4%	0.50 (0.05)	93.0%	0.97 (0.29)	0.99 (0.26)	0.02 (0.18)
Independent Correlated Differential Misclassification ^{†††}	1.00 (0.20)	95.2%	0.50 (0.05)	94.8%	0.98 (0.29)	1.01 (0.22)	
Uncorrelated Differential Misclassification ^{††††}	1.12 (0.21)	92.0%	0.51 (0.05)	93.4%	1.89 (0.36)		
Nondifferential Misclassification ^{†††††}	2.40 (0.18)	0.2%	0.48 (0.05)	91.6%	2.00 (0.33)		
Selected Model	1.00 (0.20)	95.0%	0.50 (0.05)	94.8%	0.97 (0.29)	1.00 (0.24)	

[†] ML based on (11) and (12); 500 simulations under each set of conditions with $(\beta_0, \beta_1, \beta_2, \sigma_u^2, \sigma_{u*}^2, \psi) = (0, 1, 0.5, 1, 1, 0), (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (-3, 0.05, 0.2, 0.4, 1.5, 3)$. $X_1 \sim \text{Bernoulli}(0.5)$ and $X_2 \sim N(0, 4)$. $n_m=800$, $n_v=200$.

^{††} Converged 487 times.

^{†††} Converged 500 times.

^{††††} Converged 495 times.

^{†††††} Converged 492 times.

Table 7

Simulations comparing MLEs under internal validation design for logistic-mixed regression with general outcome misclassification and two covariates ($J=4$)[†]

Model	β_1		β_2		$\hat{\sigma}_u^2$	$\hat{\sigma}_{u^*}^2$	ψ
	Mean (SD)	95% Coverage	Mean (SD)	95% Coverage	Mean (SD)	Mean (SD)	Mean (SD)
Naive	1.55 (0.08)	0	0.24 (0.02)	0	0.24 (0.07)		
Ideal	0.99 (0.10)	95.9%	0.50 (0.03)	94.7%	0.99 (0.14)		
General Misclassification ^{††}	1.02 (0.16)	96.0%	0.50 (0.05)	93.8%	0.93 (0.20)	0.06 (0.08)	-0.03 (0.11)
Independent Correlated Differential Misclassification ^{†††}	1.03 (0.16)	95.3%	0.50 (0.05)	95.5%	0.91 (0.19)	0.05 (0.08)	
Uncorrelated Differential Misclassification ^{††††}	0.98 (0.17)	96.2%	0.50 (0.05)	94.6%	0.98 (0.24)		
Nondifferential Misclassification ^{†††††}	2.26 (0.16)	0	0.47 (0.05)	86.2%	1.14 (0.24)		
Selected Model	0.98 (0.17)	96.0%	0.50 (0.05)	94.4%	0.96 (0.29)		

[†] ML based on (11) and (12); u_i^* removed; 500 simulations under each set of conditions with $(\beta_0, \beta_1, \beta_2, \sigma_u^2, \sigma_{u^*}^2, \psi) = (0, 1, 0.5, 1, 1, 0)$, $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (-3, 0.05, 0.2, 0.4, 1.5, 3)$. $X_1 \sim \text{Bernoulli}(0.5)$ and $X_2 \sim N(0, 4)$. $n_m=800$, $n_v=200$.

^{††} Converged 492 times.

^{†††} Converged 500 times.

^{††††} Converged 500 times.

^{†††††} Converged 500 times.