Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Li Tang Date

Analysis of Data with Complex Misclassification in Response or Predictor Variables by Incorporating Validation Subsampling

By

Li Tang

Doctor of Philosophy

Biostatistics

_____

Dr. Robert H. Lyles

Advisor

_____

Dr. W. Dana Flanders

Committee Member

_____

Dr. Michael J. Haber

Committee Member

_____

Dr. John J. Hanfelt

Committee Member

Accepted:

_____

Lisa A. Tedesco, Ph.D. Dean of the James T. Laney School of Graduate Studies

_____

Date

Analysis of Data with Complex Misclassification in Response or Predictor Variables by Incorporating Validation Subsampling

By

Li Tang

M.Sc. Emory University, 2006

Advisor: Robert H. Lyles, Ph.D.

An Abstract of

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2012

Abstract

Analysis of Data with Complex Misclassification in Response or Predictor Variables by Incorporating Validation Subsampling

By Li Tang

The problems of misclassification are common in epidemiological and clinical research. Misclassification may be present in either an exposure or outcome variable, or both. It is well known that the validity of analytic results (e.g., estimates of odds ratios of interest) might be questionable when no correction effort is made. Therefore, valid and accessible methods with which to deal with these issues are still in high demand.

In this dissertation, we first consider the situation when correlated binary response variables are subject to misclassification. Building upon prior work that extended McNemar's test to correct paired-data odds ratio estimation, we propose a nonlinear mixed model-based approach to adjust for potentially complex differential misclassification in correlated binary responses via internal validation sampling.

In the second topic, we shift gears toward predictor misclassification, for which we develop likelihood-based approaches based on generalized linear and generalized linear mixed models that can efficiently incorporate internal validation data in univariate and multivariate settings, respectively. We discuss the use of the approach both in the case when a baseline predictor is misclassified and when a time-dependent predictor is misclassified.

In the final topic, we elucidate extensions of well-studied methods in order to facilitate misclassification adjustment when a binary outcome and binary exposure variable are both subject to complex differential misclassification in the 2-by-2 table scenario. We develop maximum likelihood approaches to accommodate a broad range of complexity in the joint misclassification process while incorporating various types of internal validation observations. We then generalize the method to a more standard binary regression setting, allowing the incorporation of covariates both in the main health effects model of interest and in misclassification models for both the binary outcome and exposure variable. Throughout, illustrative examples are presented via detailed analyses of bacterial vaginosis and trichomoniasis data from the HIV Research Epidemiology Study (HERS).

Key Words: Differential; Misclassification; Internal Validation; Likelihood

Analysis of Data with Complex Misclassification in Response or Predictor Variables by Incorporating Validation Subsampling

By

Li Tang

M.Sc. Emory University, 2006

Advisor: Robert H. Lyles, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2012

Acknowledgment

I wish to acknowledge the support of various individuals and organizations. Particularly, I would like to express my deep gratitude to my supervisors, Dr. Robert H. Lyles for his mentorship, guidance, encouragement and support during my PhD studies. His dedications to students and academic profession are to be highly appreciated. I would also like to thank my committee members Dr. Dana Flanders, Dr. Michael Haber and Dr. John Hanfelt for reviewing my dissertation and their constructive suggestions.

I would like to specially thank faculty members, staff and fellow graduate students in the Department of Biostatistics and Bioinformatics, Rollins School of Public Health at Emory University for their continued encouragement and support.

Finally, I would like to dedicate my achievements to my parents for their love and support.

Table of Contents

List of Tables

Chapter 1 Introduction

1.1 Overview

In many clinical and epidemiologic studies, one aims to characterize the association of a health-related binary response (i.e. disease status) with predictors of interest. The ideal situation is that both response variables and predictors are measured without error. When this holds, standard statistical procedures, including generalized linear models (GLMs) for a univariate response, and generalized linear mixed models (GLMMs) or generalized estimating equations (GEE) for repeatedly measured responses, are readily available for use. However, in practice, mismeasurement in response or predictors or even both is quite common. The reason for mismeasurement lies in the fallibility in the assessment methods chosen. Sometimes less accurate methods are used for lowering the cost or for convenience. When mismeasurement occurs, employment of methods to correct for it is desirable whenever possible in order to reduce the bias in estimation.

In this dissertation, we focus on developing: (i) a likelihood-based approach that incorporates a validation sampling design in a generalized linear mixed model (GLMM) context to correct misclassification in correlated misclassified responses; (ii) a parametric misclassification-correction approach that restores validity in regression when predictors are subject to misclassification; (iii) an accessible parametric approach to correct for the bias due to misclassification in both response and predictor variables.

1.2 Misclassification in Correlated Binary Responses

Many researchers have investigated the impacts of binary response misclassification on statistical inference. It is widely known that response misclassification can lead to severe bias as well as loss in efficiency (1-4). There is broad literature on methods correcting for response misclassification, mainly in the context of ordinary logistic regression, with the use of validation data (under certain assumptions) or known misclassification probabilities (3, 5-11). In the case of generalized linear mixed models, Neuhaus (4) quantified the magnitude of the bias when the response is misclassified. He also showed that the class of generalized linear models shares a closure property when the misclassification probabilities are independent of the covariates, allowing for the development of a computationally efficient maximum likelihood (ML) algorithm. He further pointed out that the analysis that corrects for the impact from the error in the response variable leads to efficiency loss as compared to the analysis using the error-free responses.

Regarding the use of error-assessment data, more recent literature illustrates approaches to incorporate validation data into the estimation of the regression coefficients when the outcome is differentially misclassified, via the use of a Bayesian framework (12-14), nonparametric kernel methods (15), or the use of the likelihood-based methods (3). Given efficient optimization tools in standard software nowadays, Lyles et al. (16) demonstrated a computationally more accessible ML method to correct for differentially misclassified binary outcomes in ordinary logistic regression by using internal validation data.

Despite a wide range of choices on the correction methods, most attention has been focused on the case when there is no repeated measurement on the response. However, longitudinal studies have been common in practice. An efficient and computationally accessible method to adjust for differential misclassification in correlated binary outcomes is in demand. Taking the advantage of the closure property of the class of generalized linear models when responses are misclassified, Neuhaus (17) proposed a general framework to implement population-averaged (GEE) and cluster-specific (GLMM) analyses when the misclassification probabilities are known or unknown but fixed and independent of covariates. Neuhaus (17) also pointed out that the closure property still holds when the misclassification probabilities depend on covariates via a known deterministic function. He implied that when the misclassification probabilities depend on covariates via a function with unknown parameters, theoretically ML estimates can be obtained with nonlinear optimization tools, but identifiability issues may arise. Subsequently, Lyles et al (18) specifically examined a case of matched-pair 2×2 tables in a longitudinal study. When the 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. They also provided guidance for assessing cost-efficiency in terms of the study design. This likelihood-based method can be viewed as an idea closely related to an extension of conditional likelihood or GLMM.

In this dissertation, we first focus on longitudinal studies with repeatedly measured responses misclassified. We integrate and extend the previous work of Neuhaus (17) and Lyles et al (18). Neuhaus (17) has provided a general likelihood expression that can accommodate correlated binary outcome misclassification, but little guidance was given

when the misclassification was differential and when there was a need of incorporating validation data. Lyles et al (18) successfully incorporated the idea of validation data, but did not address the situation when multiple covariates need to be adjusted for. We seek to demonstrate a likelihood-based method that reliably estimates the parameters of interest in a generalized linear regression setting in the presence of differentially misclassified correlated binary outcomes, by incorporating validation data. We also aim to make the method more practically accessible by taking advantage of built-in optimization tools in standard software.

1.3 Misclassification in Predictors

Misclassification in predictor variables is a long-standing problem in statistics. It has been widely known that misclassification on exposure can potentially cause severe bias in parameter estimation, when the primary goal of a study is to characterize the association between an outcome and a pool of predictors (1, 3, 19, and 20). Further, the direction of the bias is known to rely heavily on the pattern of misclassification. Non-differential misclassification, in which the misclassification probabilities do not depend on the outcome or other covariate status, is often believed to cause a bias towards the null, when other conditions are met (3, 20-24). In contrast, when differential misclassification is present, the direction of bias is hard to predict (26).

When there is a validation sample available, misclassification probabilities can be estimated, and thus parameter estimates can be adjusted accordingly. Methods developed to incorporate validation data include likelihood-based approaches and estimating equation approaches (3, 6, 8, 26, and 27). When a gold standard measurement is not available, replicates of the error-prone measurements can be used for the correction,

assuming nondifferential misclassification (28-35). When there is no validation or replicated data available, investigators may rely on sensitivity analysis to infer the range of bias in the estimation by supplying a series of values for misclassification parameters and assuming those values are known (36-38).

Liu and Liang (33) presented a method to correct for binary predictor misclassification in generalized linear models. In their approach, they considered generalized linear models with predictors non-differentially misclassified, and demonstrated the use of quasi-likelihood for obtaining corrected estimators. The misclassification probabilities were estimated from replicates, and the number of replicates to maintain a desired efficiency was discussed. In related work, Kosinski and Flanders (39) outlined an EM algorithm that can be implemented in standard statistical software to account for nondifferential or differential exposure misclassification in regression analysis by using two imperfect measurement methods.

Although there is much literature on predictor misclassification, so far few references have provided detailed information on handling differential predictor misclassification when incorporating validation data. In this dissertation, we assume the primary objective of a study is to assess the association of a response with one or more health-related binary predictors measured with error in generalized linear models, adjusting for other covariates. We focus on developing a likelihood-based approach that can be accommodated using standard software.

1.4 Misclassification in Response and Predictor Variables in 2×2 Tables

Known as the "matrix method" in epidemiological textbooks (40, 41), an intuitive correction approach due to Barron (1) can be parameterized in terms of familiar sensitivity and specificity properties of surrogate measurements on disease and exposure status, assuming nondifferential and independent misclassification in both. When misclassification in either is negligible, Greenland (6) derived variance estimators for differential and non-differential misclassification using the "matrix method", under various validation sampling schemes. Adopting an alternative way of parameterizing with positive and negative predictive values, Marshall (8) developed another equality-based correction method, designated as the "inverse matrix method" (9). However, the original use of the "inverse matrix method" is restricted to the situation when either disease or exposure status is differentially misclassified, under which it has since been shown that Marshall's closed-form internal validation data-based corrected odds ratio estimator is also an ML estimator (42). Extensive discussions concerning efficiency of the matrix and inverse matrix methods versus the ML approach with misclassified exposure can be found in the literature (9).

Though rich literature is available for correcting misclassification in a binary variable, little guidance has been provided to deal with epidemiologic and clinical data with both response and exposure variables subject to misclassification. Therefore, we envision the practical need of developing intuitive methods for estimating odds ratios in 2 x 2 tables with a more general view, and we also sense the significance of making the correction methods computationally friendly and accessible. Greenland and Kleinbaum (26) offered an extended version of the matrix method (41) involving differential misclassification,

and this method is advantageous in terms of flexibly allowing for differential misclassification in either or both variables. Here, we seek to further extend the focus within the 2 x 2 table setting with both variables misclassified, while enhancing the ease in computation and conceptualization and allowing for flexible modeling of misclassification in both variables via internal validation data.

We first provide an ML framework allowing for flexible modeling of misclassification in both variables. We demonstrate that the ML approach can be viewed as directly connected to generalized versions of matrix and inverse matrix methods, while sharing common elements with prior work (3, 5 and 9). Compared to well-studied methods, our approach is more general in the sense of handling a richer set of misclassification patterns. To the best of our knowledge, this is also the first time that the inverse matrix method is fully generalized. We draw comparisons between different methods, and make our suggestions for analyzing data in practice. We also emphasize the advantage of utilizing an internal validation subsample in facilitating efficient estimation of corrected odds ratios. A model selection procedure readily implemented in standard statistical software is provided to practicing epidemiologists and clinicians. Throughout, the primary focus is on the point estimation of odds ratios in cross-sectional studies. However, notes will be offered on null testing and the applicability of the methods to case-control studies.

1.5 Misclassification in Response and Predictor Variables in Regression

As 2×2 tables can be viewed as special cases of regressions, a natural extension of the topic addressed in Section 1.4 would be an approach that corrects biases in coefficients when both response and predictor variables are subject to misclassification. The advantage of a regression-based correction approach is that it makes it possible to control

for covariate information to make more accurate inference. Most prior attention in regression settings has been focused on exposure misclassification (3), though some attention has been offered to response misclassification (4, 12, 17). More importantly, little guidance has been provided for analyses with complex misclassification mechanisms such as differential misclassification, when subject's covariate information has an impact on the misclassification process. Following the idea of specifying misclassification models (16), we aim to propose an approach to accommodate complex misclassification in both response and predictor variables simultaneously.

1.6 Motivating Example

The motivating examples for this dissertation are taken from the HIV Epidemiology Research Study (HERS). This is a multi-center prospective cohort study with a total of 1310 women enrolled in four U.S. cities from 1993 to 1995 (43). Among them, 871 women were HIV-infected, and 439 were not infected but at risk. During each semi-annual visit, a wealth of health-related information was collected.

The first question of interest is to assess the prevalence of bacterial vaginosis (BV) when adjusting for necessary covariates. BV was measured by two different clinical methods: the clinically-based (CLIN) and the laboratory-based (LAB). CLIN was a less accurate method that diagnoses BV by evaluating multiple clinical characteristics based on a modified Amsel's criteria (44), while LAB relies on a more sophisticated Gram-staining technique (45). The LAB method is more expensive and serves here as an arguable gold-standard method, while the CLIN method is more cost-efficient and accessible. An important feature of the HERS data is that both CLIN and LAB diagnoses

were recorded at each visit, which makes it possible for us to thoroughly evaluate the performance of the proposed approach.

A second research question to be addressed is the covariate-adjusted association of elevated vaginal PH (>4.7) with trichomoniasis. The predictor of interest that is subject to misclassification is trichomoniasis status. Trichomoniasis status was also diagnosed semiannually by two techniques, wet mount (WET) and culture testing (CULTURE). Wet mount is the most common clinical method for diagnosing trichomoniasis; however, it suffers from a low sensitivity (50). For a wet mount, a clinician visually examines a microscope slide, prepared by suspending a specimen in saline solution, for trichomonads. Culture testing is considered the gold standard, which is highly sensitive and specific with negligible error. For culture testing, trichomonads are examined from a specimen placed in a culture medium for 2-7 days. Compared to wet mount, culture is more expensive and delays diagnostic results (50).

Our third research question addresses how BV is associated with trichomoniasis. It is widely believed that BV and trichomoniasis are associated, since the presence of trichomoniasis tends to create a bacterial favorable environment, leading to a higher chance of BV incidence (51). Thus, we consider CLIN BV as an error-prone substitute (Y^*) for LAB BV (Y), while WET TRICH as an error-prone predictor (X^*) replacing CULTURE TRICH (X). In contrast to the first and second questions above, both BV and trichomoniasis are subject to misclassification error. Thus, this setting provides an available analytic example to evaluate the performance of the proposed approaches to the problems outlined in Sections 1.4 and 1.5 (i.e., with and without covariate adjustment).

Chapter 2 Regression Analysis for Differentially Misclassified Correlated Binary Responses

2.1 Methods

2.1.1 Notation

Let Y_{ij} be the true response of interest for subject i at the jth occasion, with Y_{ij}=1 if disease is present and Y_{ij}=0 otherwise. The response depends on a set of covariates $\mathbf{X}_{ij}$ =($X_{ij1},\ldots,X_{ijp}$). Suppose that the association between the two follows a generalized linear mixed model:

$$g\{\Pr(Y_{ij}=1 \mid X_{ij})\} = \beta_0 + \boldsymbol{X_{ij}}^{\boldsymbol{T}}\boldsymbol{\beta} + \boldsymbol{Z_{ij}}^{\boldsymbol{T}}\boldsymbol{U_i} \qquad (2.1),$$

in which g is an arbitrary link, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T$ is a parameter vector with p dimensions, $\mathbf{Z}_{ij}$ =($Z_{ij1},\ldots Z_{ijq}$) is a regressor vector for random effects, and $\mathbf{U_i}=(u_{i1},\ldots, u_{iq})^T$ is a subject-specific random effect vector that follows a multivariate distribution f($\mathbf{U_i}$). Conditioned on $\mathbf{U_i}$, it is assumed that responses within subject i are conditionally independent (46). With such an assumption, the likelihood of Y_{ij} can be fully specified as a function of β_0, $\boldsymbol{\beta}$, $\mathbf{U_i}$ and parameters involved in f($\mathbf{U_i}$). Often f($\mathbf{U_i}$) is assumed to be a multivariate normal distribution N(0, $\boldsymbol{\Sigma}$), which will be followed throughout.

When Y_{ij} is misclassified, instead of observing Y_{ij}, information on error-prone responses Y^*_{ij} is collected. Y^*_{ij} relates to Y_{ij} via the bridge of clinical properties of the diagnostic tools, known as sensitivity and specificity in epidemiology. For non-differential misclassification as defined in equation (2.2),

$$SE = Pr\left(Y_{ij}^* = 1 \middle| Y_{ij} = 1\right) \text{ and } SP = Pr\left(Y_{ij}^* = 0 \middle| Y_{ij} = 0\right) \qquad (2.2)$$

we assume that both sensitivity and specificity are fixed constants across all subjects and occasions, and they are independent of other information. For example, $Pr\left(Y_{ij}^* = y_{ij}^* \middle| \boldsymbol{Y_i}, \boldsymbol{C_i}\right) = Pr\left(Y_{ij}^* = y_{ij}^* \middle| Y_{ij} = y_{ij}\right)$, where $\mathbf{C_i}$ represents a set of time-dependent and/or time-independent subject-specific characteristics.

In contrast, when misclassification is differential, we allow sensitivity and specificity to vary under different conditions as in (2.3).

$$SE_{c_{ij}} = Pr\left(Y_{ij}^* = 1 \middle| Y_{ij} = 1, \boldsymbol{C_{ij}}\right) \text{ and } SP_{c_{ij}} = Pr\left(Y_{ij}^* = 0 \middle| Y_{ij} = 0, \boldsymbol{C_{ij}}\right) \qquad (2.3)$$

We assume that $\mathbf{C_{ij}}$ in eqn. (2.3) is a low-dimensional and quantifiable vector ($\mathbf{C_{ij1}}^{\mathbf{T}}$,…, $\mathbf{C_{ijq}}^{\mathbf{T}}$)$^{\mathrm{T}}$ that is not necessarily the same as $\mathbf{X_{ij}}$. $\mathbf{C_{ij}}$ may be a subset of $\mathbf{X_{ij}}$, or may not overlap with $\mathbf{X_{ij}}$ at all. The subscript in $\mathbf{C_{ij}}$ indicates that misclassification rates depend on subject-specific information. Henceforth, we assume that sensitivity and specificity for subject i at the j-th occasion only depends on the corresponding true response and covariate information at time point j, and that the misclassification processes for different occasions within the same subject are conditionally independent. In many situations, such conditional independence is a sensible assumption. More specifically, we assume that $Pr\left(Y_{ij}^* = y_{ij}^* \middle| \boldsymbol{Y_i}, \boldsymbol{C_i}\right) = Pr\left(Y_{ij}^* = y_{ij}^* \middle| Y_{ij} = y_{ij}, \boldsymbol{C_{ij}}\right)$. Nevertheless, a note is given in Section 2.1.7 regarding the case when misclassification correlates within the same subject.

For both non-differential and differential misclassification cases, SE and SP can be estimated via validation data. Otherwise, possible values of SE and SP can be supplied by users for the purpose of sensitivity analysis.

2.1.2 Validation Sampling Scheme

External validation data usually are separate and independent from a current study sample (i.e., from previous similar studies, or literature) (9). To incorporate the information from the external validation data, one must assume "transportability", i.e., that the misclassification probabilities operating in the external validation set are the same as those operating in the current study (3). This assumption is questionable and sometimes unverifiable in practice. In external studies, typically, only information on (Y_i, Y_i^*) would be available; note the lack of any 'j' subscript corresponding to time points in the main study. In most of the cases, the use of external validation data forces the fully non-differential misclassification assumption because other information is not available.

Unlike external validation sampling, internal validation involves a randomly selected proportion of the study sample, and in this subsample the true response is measured. Benefits of such a type of design include avoidance of the assumption of transportability, improved efficiency, and the flexibility of allowing for differential misclassification.

2.1.3 Non-differential Misclassification with External Validation

As mentioned in Section 2.1.2, when limited to external validation data, we usually can only consider the cases of non-differential and independent misclassification. Let n_m be the number of subjects in the main study and n_v be the number in the external validation sample. Assuming non-differentiality and independence, we can show that:

$$Pr\left(Y_{ij}^* = 1 \middle| \boldsymbol{X_{ij}}, u_i\right) = \sum_{y_{ij}=0}^{1} Pr\left(Y_{ij}^* = 1 \middle| Y_{ij} = y_{ij}\right) Pr\left(Y_{ij} = y_{ij} \middle| \boldsymbol{X_{ij}}, u_i\right)$$

$$= (1 - SP\ \ ) + (SE\ \ + SP\ \ - 1)\, Pr\left(Y_{ij} = 1 \middle| \boldsymbol{X_{ij}}, u_i\right) \quad (2.4)$$

Equation (4) has the same form as shown by Neuhaus (17). Thus, the likelihood of the main study has the following form:

$$L_m = \prod_{i=1}^{n_m} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \prod_{j=1}^{J} \{\left[(1 - SP\ \ ) + (SE\ \ + SP\ \ - 1)\, Pr\left(Y_{ij} = 1 \middle| \boldsymbol{X_{ij}}, \boldsymbol{Z_{ij}}, \boldsymbol{U_i}\right)\right]^{y_{ij}^*}$$

$$\times \left[SP\ \ - (SE\ \ + SP\ \ - 1)\, Pr\left(Y_{ij} = 1 \middle| \boldsymbol{X_{ij}}, \boldsymbol{Z_{ij}}, \boldsymbol{U_i}\right)\right]^{\left(1 - y_{ij}^*\right)} f(\boldsymbol{U_i}) d\boldsymbol{U_i}\} \quad (2.5)$$

With external validation with pairs of (Y^*_k, Y_k) (k=1,…, n_v) observed, we can derive the likelihood contribution for each pair as follows:

$$Pr(Y_k^* = y_k^*, Y_k = y_k) = Pr(Y_k^* = y_k^* | Y_k = y_k) Pr(Y_k = y_k) \quad (2.6)$$

in which the first term reflects the misclassification probability and the second term is a nuisance parameter reflecting the prevalence of Y_k=y_k in the external validation sample. Therefore, the likelihood for the external validation sample has a similar form as derived in Lyles et al (16):

$$L_v = \prod_{k=1}^{n_v} \{[SE \times Pr(Y_k = 1)]^{y_k^* y_k} \times [(1 - SE) \times Pr(Y_k = 1)]^{(1 - y_k^*) y_k}$$

$$\times [(1 - SP) \times Pr(Y_k = 0)]^{y_k^* (1 - y_k)} \times [SP \times Pr(Y_k = 0)]^{(1 - y_k^*)(1 - y_k)}\} \quad (2.7)$$

The full likelihood incorporating external validation is proportional to L_mxL_v.

2.1.4 Differential Misclassification

To model misclassification probability is a function of covariates, we introduce a secondary generalized linear model with an arbitrary link g' that may or may not be the same as the link g in equation (2.1):

$$g'\{\Pr\left(Y_{ij}^{*}=1\middle|\, Y_{ij}, \boldsymbol{C_{ij}}\right)\}=\gamma_0+\sum_{k=1}^{k=q}\gamma_k\, C_{ijk}+\gamma_{q+1}y_{ij} \qquad (2.8)$$

As noted in Section 2.1.2, we assume independent misclassification processes here, however, this type of setting allows sensitivity and specificity to differ regarding subject-specific information. It also makes the likelihood ratio (LR) test an option to select variables that have important impacts on the sensitivity and specificity, and those covariates may or may not overlap with the ones in eqn. (2.1). A likelihood ratio test can also be used to check for the assumption of overall non-differentiality (H_0: $\gamma_1=\ldots=\gamma_q=0$).

Similar to the case incorporating external validation data, the likelihood contribution from the main study can be derived accordingly:

$$L_m =$$

$$\prod_{i=1}^{n_m}\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}\sum_{j=1}^{J}\{\left[\left(1-SP_{c_{ij}}\right)+\left(SE_{c_{ij}}+SP_{c_{ij}}-1\right)\times Pr\left(Y_{ij}=1\middle|\boldsymbol{X_{ij}},\boldsymbol{Z_{ij}},\boldsymbol{U_i}\right)\right]^{y_{ij}^{*}}$$

$$\times\ \left[SP_{c_{ij}}-\left(SE_{c_{ij}}+SP_{c_{ij}}-1\right)Pr\left(Y_{ij}=1\middle|\boldsymbol{X_{ij}},\boldsymbol{Z_{ij}},\boldsymbol{U_i}\right)\right]^{\left(1-y_{ij}^{*}\right)}f(\boldsymbol{U_i})d\boldsymbol{U_i}\} \qquad (2.9)$$

The difference of eqn. (2.9) from eqn. (2.7) is that in the case of the internal validation design, misclassification probabilities can be modeled a function of covariates, which adds much more flexibility into the analysis.

Unlike the case of external validation, there may be multiple possible types of internal validation data that can be collected. For example, for pairwise correlated responses, as described in Lyles et al (18), we might consider three types of internal validation data:

(Y_{i1}^*,Y_{i2}^*,Y_{i1}) (Type I), (Y_{i1}^*,Y_{i2}^*,Y_{i2}) (Type II), and (Y_{i1}, Y_{i2}) (Type III). Type III facilitates a cost-efficiency analysis and optimal design to assess whether we need any surrogate measurements for a cost-optimal design. For studies with more than two repeatedly-measured outcomes, similar designs can be implemented without conceptual difficulty. Let n_{v1}, n_{v2} and n_{v3} be the number of subjects randomly assigned into Type I, II and III. Without loss of generality, we assume that the data is sorted with the first n_m subjects forming the main study, and the following n_{v1} subjects as the Type I internal validation subset and so on. For the case when there are two correlated binary responses for each subject, Type I subjects in the internal validation sample contribute to the likelihood via the form in eqn. (2.10).

$$L_{v1} = \prod_{i=n_m+1}^{n_m+n_{v1}} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \{[SE_{c_{i1}} \times P_{Y_{i1}}]^{y_{i1}^* y_{i1}} [(1 - SE_{c_{i1}}) \times P_{Y_{i1}}]^{(1-y_{i1}^*)y_{i1}}$$

$$\times \left[(1 - SP_{c_{i1}}) \times (1 - P_{Y_{i1}})\right]^{y_{i1}^*(1-y_{i1})} \left[SP_{c_{i1}} \times (1 - P_{Y_{i1}})\right]^{(1-y_{i1}^*)(1-y_{i1})}$$

$$\times \left[(1 - SP_{c_{i2}}) + (SE_{c_{i2}} + SP_{c_{i2}} - 1) P_{Y_{i2}}\right]^{y_{i2}^*}$$

$$\times \left[SP_{c_{i2}} - (SE_{c_{i2}} + SP_{c_{i2}} - 1) P_{Y_{i2}}\right]^{(1-y_{i2}^*)} f(\boldsymbol{U_i}) d\boldsymbol{U_i}\} \qquad (2.10)$$

where $P_{Y_{i1}} = \Pr(Y_{i1} = 1 | \boldsymbol{x_{i1}}, \boldsymbol{z_{i1}}, \boldsymbol{u_i})$ and $P_{Y_{i2}} = \Pr(Y_{i2} = 1 | \boldsymbol{x_{i2}}, \boldsymbol{z_{i2}}, \boldsymbol{u_i})$.

Similarly, for type II and III, the likelihood components are shown in equations (2.11) and (2.12).

$$L_{v2} = \prod_{i=n_m+n_{v1}+1}^{n_m+n_{v1}+n_{v2}} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \{[SE_{c_{i2}} \times P_{Y_{i2}}]^{y_{i2}^* y_{i2}} [(1 - SE_{c_{i2}}) \times P_{Y_{i2}}]^{(1-y_{i2}^*)y_{i2}}$$

$$\times \left[(1 - SP_{c_{i2}}) \times (1 - P_{Y_{i2}})\right]^{y_{i2}^*(1-y_{i2})} \left[SP_{c_{i2}} \times (1 - P_{Y_{i2}})\right]^{(1-y_{i2}^*)(1-y_{i2})}$$

$$\times \left[\left(1 - SP_{c_{i1}}\right) + \left(SE_{c_{i1}} + SP_{c_{i1}} - 1\right)P_{Y_{i1}}\right]^{y_{i1}^*}$$

$$\times \left[SP_{c_{i1}} - \left(SE_{c_{i1}} + SP_{c_{i1}} - 1\right)P_{Y_{i1}}\right]^{\left(1-y_{i1}^*\right)} f(\boldsymbol{U_i}) d\boldsymbol{U_i}\} \qquad (2.11)$$

$$L_{v3}$$

$$= \prod_{i=n_m+n_{v1}+n_{v2}+1}^{n_m+n_{v1}+n_{v2}+n_{v3}} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \{\Pr\left(Y_{i1} = y_{i1} \left| \boldsymbol{x_{i1}}, \boldsymbol{z_{i1}}, \boldsymbol{u_i}\right.\right) \Pr\left(Y_{i2} = y_{i2} \left| \boldsymbol{x_{i2}}, \boldsymbol{z_{i2}}, \boldsymbol{u_i}\right.\right) f(\boldsymbol{U_i}) d\boldsymbol{U_i}\} \qquad (2.12)$$

The full likelihood when incorporating internal validation is proportional to L_mx L_{V1}xL_{V2}xL_{V3}.

For the purpose of modeling SE/SP differentially, we introduce a secondary model (eqn. (2.8)). A careful model selection should be performed to ensure important covariates **C** are included properly. Failing to include important covariates in eqn. (2.8) may result in invalidity in estimating regression coefficients in the main model (2.1). However, this is not always true. From the likelihood of L_mxL_V with L_m and L_v defined above, if the left-out covariate in eqn. (2.8) is not in the main model (2.1), the MLE of its corresponding β coefficient in eqn. (2.1) would not be affected. Only if the left-out variable is also a predictor for the main model, the MLE of β will be invalidated. (See Section 2.3.3 for details).

2.1.5 Main-study Only and Sensitivity Analysis

Under some situations, validation data will not be available. When that is the case, sensitivity analysis is a reasonable choice. Neuhaus (17) showed that with non-differential misclassification, the likelihood-based approach can be applied to incorporate sensitivity analysis by supplying a series of possible SE and SP values into eqn. (2.4) and (2.5). He also investigated the possibility of obtaining estimates on **β** in the main model

(2.1) without supplying SE and SP values when assuming non-differentiality, but numerical issues may arise in this case, as identifiability is likely weak (3).

2.1.6 Estimation

The likelihood can be optimized after integrating out the random effects. Many numerical methods are available for the numerical integration, including adaptive Gaussian quardrature (47) and the first-order method (48). The full likelihood can then be optimized via quasi-Newton optimization, and standard errors of estimates may be obtained on the basis of the appropriate Hessian matrix. Such optimization techniques are well developed in standard software. All simulations and data examples are carried out using the NLMIXED procedure in SAS 9.2 (49) unless otherwise specified.

2.1.7 Correlation in Misclassification Processes

If repeated responses within the same individual are measured using the same defective device, any two misclassification processes for that subject may or may not correlate with each other, depending on the situation. In our motivating HERS BV example, it is sensible to assume that misclassification is independent conditioned on observed covariates. In other cases, misclassification may be correlated. Generally, if only external validation information is available, it is difficult to assess the correlation in the misclassification processes because the external validation sample may not share similar clustering information. When an internal validation sample is available, it is possible to evaluate potential correlation in the misclassification processes by including another random effect term in eqn. (2.8) to accommodate correlations in misclassification.

2.2 Simulation Studies

In this section, we describe simulation studies to assess the performance of the proposed approaches under different situations. In all cases we confine our attention to the case where g and g' in models (2.1) and (2.8) are logit links. However, they can be easily generalized to other links. Unless specified, simulations were conducted for the case when there were two repeated measurements on the response for each subject, misclassification was assumed to be independent and only a random intercept was involved in model (2.1). The random effects u_i generated in eqn. (2.1) followed an i.i.d. N(0,1).

2.2.1 Non-differential Misclassification

Table 2.1 summarizes the simulation results comparing the performance of MLEs based on eqn.s (2.5) and (2.7) with different types of internal validation designs. In each case, data were generated via model (2.1) with a continuous normal covariate X with mean 0 and variance 4. True values for β_0 and β_1 were 0 and 1.0. True sensitivity and specificity were 0.70 and 0.85 respectively. For each simulated sample there were 500 main study observations and either 150 external with (Y_i, Y_i^*) or internal validation observations with (Y_{ij},Y_{ij}^*) observed. We conducted simulations for the case of 2 correlated responses and for the case of 5 correlated responses. A total of 1000 simulations were performed for each. Five models were fitted. The ideal analysis was regression with the true response as the outcome. The "naïve" analysis regressing with the error-prone outcome, two models incorporated either external or internal validation, and the final model only used the main study observations.

Table 2.1 Results of simulations comparing MLEs under external/internal validation design and under main study only for logistic-mixed regression with non-differential outcome misclassification and a single continuous predictor X.†

2 Correlated Responses								
Model	$\hat{\beta}_0$		$\hat{\beta}_1$		$\widehat{\sigma_u^2}$	$\widehat{SE}$	$\widehat{SP}$	Successful Optimization
	Mean (SD)	95% Coverag	Mean (SD)	95% Covera	Mean (SD)	Mean (SD)	Mean (SD)	
Naive	-0.35 (0.07)	0.002	0.37 (0.04)	0	0.21 (0.24)	NA	NA	1000
Ideal	0.004 (0.10)	0.96	1.00 (0.09)	0.95	1.02 (0.44)	NA	NA	1000
Main+External	-0.001 (0.31)	0.97	1.01 (0.26)	0.95	0.98 (0.86)	0.70 (0.03)	0.85 (0.03)	994
Main+Internal	0.005 (0.15)	0.96	1.01 (0.14)	0.95	1.03 (0.64)	0.70 (0.03)	0.85 (0.02)	1000
Main Study Only	-0.004 (0.48)	0.95	1.02 (0.50)	0.89	0.91 (0.84)	0.73 (0.09)	0.87 (0.07)	939
4 Correlated Responses								
Model	$\hat{\beta}_0$		$\hat{\beta}_1$		$\widehat{\sigma_u^2}$	$\widehat{SE}$	$\widehat{SP}$	Successful Optimization
	Mean (SD)	95% Coverag	Mean (SD)	95% Covera	Mean (SD)	Mean (SD)	Mean (SD)	
Naive	-0.35 (0.05)	0	0.37 (0.03)	0	0.18 (0.08)	NA	NA	1000
True	-0.004 (0.07)	0.95	1.00 (0.06)	0.95	0.99 (0.19)	NA	NA	1000
Main+External	0.004 (0.22)	0.95	1.02 (0.20)	0.93	0.96 (0.50)	0.70 (0.03)	0.85 (0.02)	998
Main+Internal	-0.002 (0.11)	0.96	1.00 (0.08)	0.95	0.98 (0.28)	0.70 (0.02)	0.85 (0.01)	1000
Main Study Only	-0.002 (0.28)	0.96	1.00 (0.30)	0.94	0.93 (0.54)	0.71 (0.05)	0.86 (0.04)	977

† ML based on eqn. (2.5) and (2.7). 1000 simulations of each set of condition. $\beta_0 = 0$, $\beta_1 = 1$, σ^2_u=1, SE=0.7, SP=0.85, n_m=500, n_v=150, and X normally distributed with mean 0 and variance 1 in each case.

The upper half of Table 2.1 represents the results when J=2, and the lower half is the results for J=4. The naïve analysis with the error-prone response as the outcome produces drastically biased results with β_1 greatly attenuated in both cases. With validation data incorporated, (β_0, β_1) are reliably estimated with excellent 95% confidence interval coverage. As compared to the analysis with external validation, "main+internal" is more numerically stable and more efficient. Estimates of sensitivity and specificity are similar in models with external and internal validation data, regarding the point estimate and precision. Predictably, more numerical problems are observed for the main-study only analysis without any validation data. Although the validity appears preserved in this case, efficiency is lost noticeably. Unsurprisingly, with more repeated measures in each subject, the precision in estimating β and SE/SP generally improves.

2.2.2 Differential Misclassification

In Table 2.2, we examine the performance of the MLE based on the proposed approach in a hypothetical longitudinal study with two time points, when the misclassification is differential. The main model (2.13) includes a binary covariate X_1 following Bernoulli (1,0.5) and a continuous covariate X_2 following N(0,4). True values of the β's are (0, 0.5, 0.5).

$$logit\left[\Pr\left(Y_{ij}=1\middle|X_{ij},u_i\right)\right]=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+u_i \qquad (2.13)$$

A secondary logistic model (2.14) was used to allow misclassification probabilities to depend on individuals' covariate information.

$$logit[\Pr\left(Y_{ij}^{*}=1\middle|Y_{ij},T,X_{ij}\right)]=\gamma_0+\gamma_1 t_{ij}+\gamma_2 x_{1i}+\gamma_3 t_{ij}\times x_{1i}+\gamma_4 y_{ij} \qquad (2.14)$$

The true SE/SP model contains an index variable t to indicate whether it is the first time point or the second time point, X_1 as in model (2.14), and their interaction term. The true values of (γ_0, γ_1,γ_2,γ_3,γ_4) were (-2, 0.5, -1.6, -0.5, 4). A total of 1000 sets of data were generated. For each simulated sample, 1000 subjects were generated, of which 500 constituted the main study observations, and the other 167, 167 and 166 were randomly assigned as the type I, II and III internal validation observations as defined in Section 2.1.4. We conducted the naïve analysis, the ideal analysis and the analyses incorporating internal validation assuming either differentiality as dedicated by model (2.14) or assuming non-differentiality.

Naïve analysis biased the estimate for β_1 in the wrong direction and attenuated the estimate for β_2. Note that only the binary covariate X_1 affects the SE/SP model. When wrongly assuming non-differentiality, $\hat{\beta}_2$ is still approximately unbiased but $\hat{\beta}_1$ is not. As discussed in Section 2.1.4, the invalidity of the estimate for β_1 relates to the fact that X_1 is an important covariate in the SE/SP model, so failure to include X_1 in the SE/SP model by wrongly assuming non-differentiality leads to invalid estimation of the corresponding coefficient. Only when the SE/SP model is specified correctly to handle the differentiality, do we find evidence supporting both estimates for β are valid. Also note from Table 2.2 that loss of efficiency is observed when estimating β_1, but efficiency loss is not noticeable when estimating β_2. There is also a small bias downward in $\widehat{\sigma_u^2}$. It is possible to perform a hypothesis test to check whether it is necessary to adjust for differential misclassification (i.e. in this case, H_0: γ_1=γ_2 =γ_3 =0). If H_0 is not rejected in practice, an easier procedure only adjusting for non-differential misclassification can be implemented. Nevertheless, for validity purposes, it may be safer to employ a more general SE/SP model.

Table 2.2 Results of simulations comparing MLEs under internal validation design for logistic-mixed regression with differential outcome misclassification and two covariates. J=2. †

Model	$\hat{\beta}_1$		$\hat{\beta}_2$		$\widehat{\sigma_u^2}$	Successful Optimization
	Mean(SD)	95% Coverage	Mean(SD)	95% Coverage	Mean(SD)	
Naive	-0.77 (0.16)	0	0.29 (0.04)	0.003	0.35 (0.26)	1000
Ideal	0.51 (0.13)	0.94	0.51 (0.06)	0.94	1.00 (0.38)	1000
Main+Internal /Differential	0.49 (0.18)	0.96	0.50 (0.05)	0.94	0.88 (0.35)	1000
Main+Internal /Non-	-0.31 (0.16)	0	0.49 (0.06)	0.94	0.94 (0.37)	1000

† ML based on eqn.s. (2.13) and (2.14); 1000 simulations under each set of conditions with (β_0, β_1, β_2, σ^2_u)= (0, 0.5, 0.5,1), (γ_0, γ_1,γ_2,γ_3,γ_4)=(-2, 0.5, -1.6, -0.5, 4). X_1 ~Bernoulli(1,0.5) and X_2 ~ N(0,4). n_m=500, n_{v1}=167, n_{v2}=167, n_{v3}=167.

2.2.3 Importance of Correctly Specifying SE/SP Model

As mentioned and discussed in section 2.1.4 and Section 2.2.2, failure to include an important covariate in the SE/SP model may lead to an invalid estimate. We further examined the impact of misspecifying the SE/SP model on the estimation. Data were generated via a main model (2.13) and an SE/SP model as eqn. (2.15):

$$logit[\Pr(Y_{ij}^* = 1 | Y_{ij}, T, X_{1i})] = \gamma_0 + \gamma_1 t_{ij} + \gamma_2 x_{1i} + \gamma_3 y_{ij} \qquad (2.15).$$

X_1 and X_2 were defined as in Section 2.2.2. True values of β's were (0, 0.5, 0.5), and (γ_0, γ_1, γ_2,γ_3)=(-2, 0.5, -1.6, 4). Note that the main model and the SE/SP model share a common covariate X_1, and time index covariate is only in the SE/SP model. ML

estimation was performed in two ways: i) X_1 was omitted from the SE/SP model specified; ii) the time index t was excluded in the SE/SP model specified. A total of 1000 simulations were conducted. In each sample, 500 were main study observations, and the other 167, 167 and 166 were randomly assigned into the type I, II and III internal validation sets as described in section 2.1.4.

Table 2.3 Results from simulation study assessing effects of omitted predictor in SE/SP model [Equation (2.15)] †

Parameter	Correct SE/SP model	X_1 omitted from SE/SP model	t omitted from SE/SP model
	Mean estimate (std. deviation)	Mean estimate (std. deviation)	Mean estimate (std. deviation)
β_1	0.48	-0.20	0.48
β_2	0.49	0.49	0.49

† ML based on eqn.s. (2.13) and (2.15); 1000 simulations under each set of conditions. 1000 simulations under each set of conditions with $(\beta_0, \beta_1, \beta_2, \sigma^2_u)$= (0, 0.5, 0.5, 1), $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$=(-2, 0.5, -1.6, 4). X_1 ~ber(1,0.5) and X_2 ~ N(0,4). n_m=500, n_{v1}=167, n_{v2}=167, n_{v3}=167.

Table 2.3 summarizes the results. Similar to the findings in section 2.2.2, when a covariate is important in both the main model and the SE/SP model, omitting it will cause invalid estimation of the main model coefficient for this covariate. In this case, X_1 has impact on the main model as well as the SE/SP model. Failing to include X_1 when specifying the SE/SP model in ML analysis thus makes $\hat{\beta}_1$ invalid. In contrast, when a covariate is only in the SE/SP model, omitting it does not affect the validity of

coefficients in the main model. From Table 2.3, incorrectly missing the time index t from the SE/SP model during the ML analysis, $\hat{\beta}_1$ and $\hat{\beta}_2$ are both approximately unbiased. In practice, care should be taken when specifying the SE/SP model. We recommend a careful model selection in practice to help ensure no important covariates are left out; note that this issue highlights the benefits of internal validation.

2.2.4 A Note About Correlated Misclassification

As mentioned in Section 2.1.6, misclassification may be correlated within the same cluster or subject, even conditioned on covariates. If that is the case, an option is to introduce another random effect u_i^* in the SE/SP model. For example, we consider simulations based on the following model:

$$logit[\Pr(Y_{ij}^* = 1 | Y_{ij}, T, X_{ij})] = \gamma_0 + \gamma_1 t_{ij} + \gamma_2 x_{1i} + \gamma_3 t_{ij} \times x_{1i} + \gamma_4 y_{ij} + u_i^* \qquad (2.16)$$

We assume the primary model of interest is the same as defined in eqn. (2.13) with a binary covariate X_1 following Bernoulli(1,0.5) and a continuous covariate X_2 following N(0,4), and we assume that u_i and u_i^* are independent. True values of β's were (0, 0.5, 0.5), and (γ_0, γ_1,γ_2,γ_3,γ_4) = (-2, 0.5, -1.6, -0.5, 4). We examined three scenarios when $\sigma_{u^*}^2$ took the value of 0.25, 1 and 2.25. A total of 500 simulations were conducted. In each sample, 500 were in the main study observations, and other 167, 167 and 166 were randomly assigned into the type I, II and III internal validation as described in section 2.1.4.

Table 2.4 summarizes the results. It is noticed that when correlation is present in misclassification, ignoring the correlation and modeling the misclassification process independently does not seem to bias the estimates of the main parameters of interest.

However, by correctly specifying the correlation in the misclassification model, efficiency is slightly improved and this trend is more obvious when the variance of u_i^* increases.

This simulation study suggests that assuming independence in misclassification when misclassification is correlated preserves validity in estimating parameters in the main model, with efficiency loss as a trade-off. Considering extra computational complications when modeling correlated misclassification processes, we suggest proceeding with the independence assumption unless the correlation is suspected to be very strong. Of course, users can always fit the model assuming independence or not, and compare the results and make conclusions based on more sensible results.

Table 2.4 Results of simulations comparing MLEs under internal validation design assuming when misclassification is correlated††.

$\sigma^2_{u^*}$=0.25					
Model	$\hat{\beta}_1$		$\hat{\beta}_2$		Successful Optimization
	Mean(SD)	95% Coverage	Mean(SD)	95% Coverage	
Naive	-0.79 (0.17)	0	0.29 (0.04)	0	500
Ideal	0.49 (0.19)	0.95	0.51 (0.05)	0.95	500
Main+Internal /Independence†	0.48 (0.19)	0.94	0.49 (0.06)	0.94	500
Main+Internal /Correlated*	0.49 (0.19)	0.95	0.49 (0.05)	0.95	500
$\sigma^2_{u^*}$=1					
Model	$\hat{\beta}_1$		$\hat{\beta}_2$		Successful Optimization
	Mean(SD)	95% Coverage	Mean(SD)	95% Coverage	
Naive	-0.83 (0.18)	0	0.28 (0.04)	0..01	500
Ideal	0.50 (0.19)	0.95	0.51 (0.06)	0.93	500
Main+Internal /Independence†	0.50 (0.19)	0.95	0.51 (0.06)	0.93	500
Main+Internal /Correlated*	0.49 (0.19)	0.95	0.50 (0.05)	0.94	500
$\sigma^2_{u^*}$=2.25					
Model	$\hat{\beta}_1$		$\hat{\beta}_2$		Successful Optimization
	Mean(SD)	95% Coverage	Mean(SD)	95% Coverage	
Naive	-0.90 (0.18)	0	0.28 (0.04)	0	500
Ideal	0.49 (0.17)	0.97	0.50 (0.05)	0.96	500
Main+Internal /Independence†	0.50 (0.20)	0.96	0.52 (0.06)	0.96	500
Main+Internal /Correlated*	0.49 (0.19)	0.96	0.51 (0.05)	0.96	500

† ML based on eqn. (2.5) and (2.7). * ML based on eqn. (2.5) and (2.16). †† Data generated based on eqn. (2.5) and (2.16). 500 simulations of each set of condition with $(\beta_0, \beta_1, \beta_2)= (0, 0.5, 0.5)$, $(\gamma_0, \gamma_1,\gamma_2,\gamma_3,\gamma_4)=(-2, 0.5, -1.6, -0.5, 4)$. X_1 ~Bernoulli(1,0.5) and X_2 ~ N(0,4). u_i, u_i^* independently follow N(0,1). n_m=500, n_{v1}=167, n_{v2}=167, n_{v3}=167. J=2.

2.3. Example

2.3.1 HERS Example

The HERS example was described in Section 1.4. We considered bacterial vaginosis (BV) as a response of interest for illustrative purposes, and BV status may be misclassified when diagnosed by the CLIN method. As aforementioned, an important feature of HERS data is that both CLIN and LAB diagnoses were recorded at each visit, which makes it possible for us to evaluate the performance of the validation design-based ML approach.

2.3.2 Example 1: Pairwise No-covariate case

Lyles et al (18) proposed an extended McNemar's test to compute paired-data odds ratio estimates for a 2×2 table with correlated binary responses misclassified, without adjusting for any other covariates. Our ML approach is equivalent to the extended McNemar's approach when specifying covariates in eqn.s (2.1) and (2.8) appropriately. As in reference (18), we consider the BV prevalence change from visit 1 to visit 4 among 565 black women. They were randomly assigned to the main study (280 women), and Type I (95 women), II (95 women) and III (95 women) internal validation sets. A paired-data odds ratio in a McNemar's test is defined as $OR = \frac{n_{D\bar{D}}}{n_{\bar{D}D}}$ (26), which is the ratio of probabilities with discordant responses at two time points.

In order to estimate the crude paired-data odds ratio, we consider a GLMM main model

$$logit\left[\Pr\left(Y_{ij} = 1 \middle| t_{ij}\right)\right] = \beta_0 + \beta_1 t_{ij} + u_i \qquad (2.17)$$

in which t_{ij}=0 (visit 4) or 1 (visit 1). The model characterizing the differential sensitivity and specificity is

$$\left[\Pr\left(Y_{ij}^{*}=1\middle|y_{ij},t_{ij},u_{i}\right)\right]=\gamma_{0}+\gamma_{1}t_{ij}+\gamma_{2}y_{ij}+\gamma_{3}t_{ij}y_{ij} \qquad (2.18)\,,$$

allowing for sensitivity/specificity to vary from visit 1 to visit 4. As shown in (18), the sensitivity/specificity of the BV status test appeared to change among black women in the HERS sample from visit 1 to visit 4. Without correcting for misclassification, the naïve analysis based on the error-prone CLIN method produced a paired-data $\widehat{OR}$ of 2.31 (95% CI=(1.49, 3.57)). However, the analysis based on the "gold-standard" LAB method gave an OR of 1.17 (95% CI=(0.90, 1.52)). The odds ratio estimate obtained without the appropriate correction was biased away from the null and led to a different hypothesis testing conclusion than the one produced by the gold-standard model.

Table 2.5 Change in BV prevalence as measured by paired data OT estimates for black women between HERS visits 1 and 4.

Results reported in reference (18)			
	ln(OR)(SE)	OR(95% CI)	p-value
LAB result	0.15 (0.13)	1.17 (0.90,1.52)	0.25
Corrected result	0.39 (0.26)	1.47 (0.89,2.43)	0.12
Results from proposed random effects model*			
	ln(OR)(SE)	OR(95% CI)	p-value
LAB result	0.15 (0.16)	1.17 (0.84,1.61)	0.35
Corrected result	0.38 (0.32)	1.46 (0.78, 2.74)	0.24

* ML based on eqn.s (2.16) and (2.17). $\widehat{\sigma_u^2}$=3.81. n_m=280, $n_{v1}=n_{v2}=n_{v3}$=95.

By performing the correction using the proposed nonlinear mixed model likelihood-based approach, the odds ratio estimate after adjusting for misclassification was 1.46 (95% CI=(0.78, 2.74)), consistent with the result reported in (18). The point estimate after the correction was close to the point estimate from the analysis using the gold standard, with some loss in efficiency (Table 2.5). The estimated sensitivity and specificity at visit 1 and visit 4 are reported in Table 2.6. At visit 4, the sensitivity of the BV status test decreased markedly from visit 1 (from 0.69 to 0.49), while the specificities at the two visits were similar (Table 2.5). This finding is consistent with the findings in reference (18), although the scientific reason for the drop in the sensitivity is not clear.

Table 2.6 SE and SP estimates. Test for equal SE/SP (nondifferentiality) assumption (H_0: $\gamma_1=\gamma_3=0$; $\chi^2=12.5$,df=2,p=0.002)

SE/SP estimates reported in reference (18)			
SE_1	SP_1	SE_4	SP_4
0.69	0.92	0.49	0.92
SE/SP estimates based on proposed random effect model*			
SE_1	SP_1	SE_4	SP_4
0.68	0.91	0.45	0.89

* ML based on eqn.s (2.16) and (2.17). $n_m=280$, $n_{v1}=n_{v2}=n_{v3}=95$.

The ML approach presented provided results consistent with those from the extended Mcnemar's approach regarding the point estimate. A slight difference in precision is observed, with slightly more variability in the proposed ML approach. The reason may be that different numerical methods are used in the two approaches in obtaining standard errors. Nevertheless, this example shows that the proposed method can serve as a valuable substitute for the case of the 2×2 paired data table. More importantly, it can also accommodate covariates/confounders that the extended McNemar's test cannot (see Example 2).

2.3.3 Example 2: Pairwise Covariate-adjusted case

In the second example, we use black, white and Hispanic patients from the 1st and the 4th semi-annual visit. In total, there are 870 patients aged greater than or equal to 25 years old at enrollment with both CLIN and LAB BV as well as all risk factors measured for

those two visits. Information collected regarding potential covariates associated with the BV status includes 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 the Caucasians after performing 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%). At the 1st semi-annual HERS visit, the estimated BV prevalence was 33.8% by the CLIN method and 46.3% by the LAB method. At the 4th visit, a crude BV prevalence estimate from the CLIN method was 25.2%, and the estimate by the LAB method was 40.9%. Thus, the CLIN method seems to underestimate the prevalence of BV in the sample.

The crude sensitivity and specificity estimates in "via sexual contact" HIV risk group are 0.46 and 0.94 respectively, combining data from two visits. However, in the HIV risk group via IDU, sensitivity and specificity are estimated to be 0.63 and 0.88. Specificity seems to be higher in the risk group with sexual contact, but the sensitivity is higher in the IDU group. In the HIV negative group, sensitivity and specificity estimates are 0.62 and 0.90, while in the HIV positive group, crude sensitivity and specificity are 0.53 and 0.92, combined across visits. Thus, the sensitivity for the CLIN method is higher in the HIV negative group. Sensitivity and specificity also change over time. At visit 1, the crude sensitivity estimate is 0.59, and it drops to 0.42 at visit 4. In contrast, the specificity increases from 0.88 to 0.94. All this indicates that a complex differential misclassification process exists in this sample.

Let Y_{ij} denote BV status for subject i at the jth time point, where j=1,4. After preliminary model selection, we assume Y_{ij} follows a logistic model as in eqn. (2.19):

$$logit\left[\Pr\left(Y_{ij}=1\middle|\boldsymbol{X}_{ij},u_i\right)\right]=\beta_0+\beta_1 hivpos+\beta_2 Age+\beta_3 Riskgrp+\beta_4 Race+u_i \qquad (2.19)$$

Because the surrogate CLIN BV measurements were obtained independently at each semi-annual visit, misclassification in repeated BV status is assumed independent, conditioned on the covariates in (2.19). A random subsample of 200 patients were selected, and we randomly assigned 100 of them to Type I and 100 to Type II internal validation sets as described in Section 2.1.4. We fitted three types of models here. The first model is the one regressing with the BV status measured by the gold-standard (LAB method) as the response. This provides the ideal result that we can compare results from other models against. The second "naïve" model is the one with the BV status measured by the error-prone CLIN method as the response. The third type of model is fitted by using the ML approach proposed, assuming either non-differentiality or differentiality as dedicated by (2.19). For the naïve model, we fitted eqn. (2.20) with Y_{ij}^{*} replacing Y_{ij}:

$$logit\left[\Pr\left(Y_{ij}{}^{*}=1\middle|\boldsymbol{X}_{ij},u_i\right)\right]=\beta_0+\beta_1 hivpos+\beta_2 Age+\beta_3 Riskgrp+\beta_4 Race+u_i \qquad (2.20).$$

For the joint model to adjust for misclassification via ML, we fitted equations (2.19) and (2.21) simultaneously, with the latter allowing for differential misclassification.

$$logit\left[\Pr\left(Y_{ij}^{*}=1\middle|Y_{ij},\boldsymbol{X}_{ij}\right)\right]$$
$$=\gamma_0+\gamma_1 hivpos+\gamma_2 Riskgrp+\gamma_3 agegtmed+\gamma_4 race+\gamma_5 t+\gamma_6 y_{ij} \qquad (2.21)$$

When assuming non-differential misclassification, we remove all the covariates except y in eqn. (2.21). We note that although formal selection of the misclassification model is

not our primary focus, model (2.21) is supported by our univariate preliminary investigations of the misclassification process.

Table 2.7A and 2.7B summarizes the fit of all models. The error-prone model and the gold-standard model differ mainly in the magnitude of the estimated OR for HIV risk group (1.63 for the ideal analysis, 2.79 for the naïve model) and in the directionality of HIV status (1.04 and non-significant for the ideal analysis, 0.70 and significant for the naïve model). This implies the potential benefit for adjusting for outcome misclassification.

Table 2.7 Change in BV prevalence for women between HERS visits 1 and 4 with covariate adjustment (Ideal and Naïve Analysis)

Ideal Analysis[a]			
Variable	$\hat{\beta}$(SE)	Estimated OR	P-value
HIV Status	0.02(0.14)	1.04	0.88
Age	-0.05(0.01)	0.95	<0.0001
Risk Group	0.48(0.14)	1.63	0.0005
Race	0.92(0.14)	2.53	<0.0001
Naïve Analysis[b]			
Variable	$\hat{\beta}$(SE)	Estimated OR	P-value
HIV Status	-0.34(0.15)	0.70	0.03
Age	-0.06(0.01)	0.94	<0.0001
Risk Group	0.98(0.15)	2.79	<0.0001
Race	1.05(0.16)	3.01	<0.0001

a. ML based on eqn. (2.19).

b. ML based on eqn. (2.20).

Table 2.8 Change in BV prevalence for women between HERS visits 1 and 4 with covariates adjusted (Main+Internal Validation Analysis).

Main+Internal, Differential*			
Variable	$\hat{\beta}$(SE)	Estimated OR	P-value
HIV Status	0.26(0.39)	1.30	0.50
Age	-0.05(0.03)	0.96	0.09
Risk Group	0.46(0.42)	1.58	0.28
Race	0.93(0.40)	2.54	0.02
Main+internal, non-differential†			
Variable	$\hat{\beta}$(SE)	Estimated OR	P-value
HIV Status	-0.45(0.25)	0.64	0.08
Age	-0.07(0.02)	0.93	0.0005
Risk Group	1.34(0.27)	3.81	<0.0001
Race	1.48(0.28)	4.39	<0.0001

* SE and SP assumed to vary with the binary variables dichotomized age, HIV risk group, HIV status, race and index for time point (Visit 4 as the reference level) via model (2.21). n_m=670, n_{v1}=n_{v2} =100. $\widehat{\sigma_u}^2 = 3.40$.

† No covariates affecting SE and SP; this assumption is not supported by the data (chi-sq=47.2, P<0.0001). $\widehat{\sigma_u}^2 = 2.74$.

Table 2.9 Change in BV prevalence for women between HERS visits 1 and 4 with covariates adjusted (Main+Internal Validation Analysis with correlated misclassification).

Main+Internal, Differential, Correlated Misclassification*			
Variable	$\hat{\beta}$(SE)	Estimated OR	P-value
HIV Status	0.02(0.25)	1.02	0.93
Age	-0.04(0.02)	0.97	0.11
Risk Group	0.52(0.26)	1.68	0.05
Race	0.87(0.34)	2.39	0.01
Main+Internal, Non-Differential, Correlated Misclassification*			
Variable	$\hat{\beta}$(SE)	Estimated OR	P-value
HIV Status	-0.37(0.24)	0.69	0.12
Age	-0.06(0.24)	0.94	0.002
Risk Group	1.17(0.25)	3.23	<0.0001
Race	1.33(0.27)	3.80	<0.0001

* SE and SP assumed to vary with the binary variables dichotomized age, HIV risk group, HIV status, race and index for time point (Visit 4 as the reference level) via model (2.21). n_m=670, n_{v1}=n_{v2} =100. $\widehat{\sigma_u}^2 = 1.32$. $\widehat{\sigma_{u^*}}^2 = 3.43$.

† No covariates affecting SE and SP; this assumption is not supported by the data (chi-sq=49.9, P<0.0001). $\widehat{\sigma_u}^2 = 1.42$ $\widehat{\sigma_{u^*}}^2 = 2.39$.

Table 2.8 shows the results when assuming non-differentiality and allowing for differentiality with the independent misclassification assumption, while utilizing the main/internal validation likelihood. When assuming non-differentiality, we are forcing the condition that $\gamma_1=\gamma_2=\gamma_3=\gamma_4=\gamma_5=0$ in eqn. (2.20). This assumption is rejected by a likelihood-ratio test ($\chi^2=47$, $P<0.0001$); therefore, differential misclassification is more plausible in this case. When allowing for differential misclassification probabilities, the interpretations are similar to those when fitting the gold-standard model, with regard to the magnitudes and directionality of the estimated ORs. In contrast, when wrongly assuming non-differentiality, the results differ markedly from the analysis allowing differential misclassification and the gold-standard analysis, regarding the point estimates. When mistakenly assuming non-differentiality, interestingly, the estimated OR for HIV status is in the same direction as for the naïve analysis. This indicates that it is important to model SE/SP differentially when non-differentiality is rejected. Sensitivity and specificity are found to be significantly associated with HIV status, risk group, time index, dichotomized age and race based on the joint NL analyses. Sensitivity tends to be higher and specificity tends to be lower in younger patients ($\widehat{\gamma_3}$=-0.50, p=0.02), black women ($\widehat{\gamma_4}$=0.73, p=0.006), at visit 1 ($\widehat{\gamma_5}$=0.68, p<0.0001), HIV negative patients ($\widehat{\gamma_1}$=-0.66, p=0.02) and patients via IDU ($\widehat{\gamma_2}$=0.96, p=0.0006). This finding is consistent with the crude estimates and references (16) and (18).

We also relaxed the independence assumption on the misclassification process, and introduced a random effect in the SE/SP model. The random effects u_i^* accommodate within-subject correlations in misclassification across visits, and are assumed to be independent from the u_i's in eqn. (2.19). Table 2.9 summarizes the results when allowing

for correlated misclassification under the assumption of differentiality and non-differentiality, based on the following generalization of the model (2.21):

$$logit\left[\Pr\left(Y_{ij}^{*}=1\middle|Y_{ij},\boldsymbol{X}_{ij}\right)\right]$$
$$=\gamma_0+\gamma_1 hivpos+\gamma_2 Riskgrp+\gamma_3 agegtmed+\gamma_4 race+\gamma_5 t+\gamma_6 y_{ij}+u_i^{*} \quad (2.22)$$

With the same main study and internal validation subsample, the estimates are very similar to those obtained when assuming independence, but with improved efficiency (smaller standard errors), with HIV risk group marginally significant (p=0.05). The non-differentiality assumption is still strongly rejected (χ^2=49.9, P<0.0001).

2.3.4 Example 3: Longitudinal Analysis with >2 Time Points

In the third example, we consider 706 patients older than 25 at the time of enrollment from the 1st through the 4th semi-annual visits. The median age at enrolment was 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%). At the 1st semi-annual visit, the estimated BV prevalence was 34.3% by the CLIN method and 46.7% by the LAB method. At the 2nd visit, a crude BV prevalence estimate from the CLIN method was 32.3%, and the estimate by the LAB method was 43.6%. The prevalence was 29.5% by CLIN method and 42.8% by LAB method at the 3rd visit, and changed to 25.9% and 41.2% at the 4th visit. The prevalence of BV had a general decreasing pattern over time.

We consider three types of models here: the ideal analysis with the gold-standard (LAB method) as the response, the naïve model with the BV status measured by the error-prone CLIN method as the response, and models fitted by using the approach proposed, assuming either non-differentiality or differentiality. On the basis of

preliminary model selection, for the "gold-standard" model, we fit

$$logit\left[\Pr\left(Y_{ij}=1\middle|\boldsymbol{X}_{ij},u_i\right)\right]=\beta_0+\beta_1 hivpos+\beta_2 Age+\beta_3 Riskgrp+\beta_4 Race+u_i \quad (2.23).$$

For the naïve model, we fit equation (2.24) instead:

$$logit\left[\Pr\left(Y_{ij}^{*}=1\middle|\boldsymbol{X_{ij}},u_i\right)\right]=\beta_0+\beta_1 hivpos+\beta_2 Age+\beta_3 Riskgrp+\beta_4 Race+u_i \quad (2.24).$$

For the model to adjust for differential misclassification, we fit equation (2.23) and (2.25):

$$logit\left[\Pr\left(Y_{ij}^{*}=1\middle|Y_{ij},\boldsymbol{X}_{ij}\right)\right]$$

$$=\gamma_0+\gamma_1 hivpos+\gamma_2 Riskgrp+\gamma_3 agegtmed+\gamma_4 race+\gamma_5 vst2+\gamma_6 vst3+\gamma_7 vst4+\gamma_8 y_{ij} \quad (2.25)$$

In eqn. (2.25), we introduced (0,1) indicator variables vst2, vst3 and vst4, using visit1 as the reference level. When assuming non-differential misclassification, we removed all the covariates except y in equation (2.25). We evenly assigned 240 patients randomly chosen into one of five internal validation types: (Y_{i1}^{*}, Y_{i2}^{*}, Y_{i3}^{*}, Y_{i4}^{*}, Y_{i1}) (Type I), (Y_{i1}^{*}, Y_{i2}^{*}, Y_{i3}^{*}, Y_{i4}^{*}, Y_{i2}) (Type II), (Y_{i1}^{*}, Y_{i2}^{*}, Y_{i3}^{*}, Y_{i4}^{*}, Y_{i3}) (Type III), and (Y_{i1}^{*}, Y_{i2}^{*}, Y_{i3}^{*}, Y_{i4}^{*}, Y_{i4}) (Type IV). The remaining 466 patients were in the main study, and they contributed to observations on Y^{*} not Y.

Table 2.10 summarizes the fit of all models. In general, the results are similar to those in Section 2.3.2. The error-prone model had the coefficient for the HIV status in the opposite direction as that in the gold-standard model (with estimates ORs of 1.04 for the gold-standard model, 0.75 for the error-prone model), and also had an inflated OR estimate for HIV risk group (1.98 for the gold-standard model, 3.08 for the error-prone model). With "main+internal" correction by allowing for differentiality, the results are similar to those from the ideal analysis, though predictably with more variability in the

estimates than in the ideal analysis. Sensitivity and specificity are found to associate with each of race, HIV status, HIV risk group, time index and age significantly based on the fit of model (2.25) in the joint ML analyses. Sensitivity tends to be higher in blacks, patients at risk via IDU, HIV negative patients, younger patients, and at visit 1. This finding is consistent with Section 2.3.3.2.When assuming the erroneous non-differentiality (χ^2=46.7, P<0.0001), the estimate for HIV status is similar to the estimate from the naïve analysis, suggesting the need of adjusting for differential misclassification.

As in section 2.3.3.2, we examined the need of accounting for correlations in misclassification across time. Table 2.12 summarizes the results. In general, with the correlations taken into account, the regression coefficient estimates in the primary model are very close to those obtained when assuming independence. The non-differentiality assumption is still rejected in this case (P<0.0001), again strongly suggesting that we should adjust for differential misclassification.

Table 2.10 Change in BV prevalence for women from HERS visits 1 through visit 4 with covariates adjusted (Ideal and Naive Analysis).

Ideal Analysis[a]			
Variable	$\hat{\beta}$(SE)	Estimated OR (95% CI)	P-value
HIV Status	0.04(0.17)	1.04	0.82
Age	-0.06(0.01)	0.94	<0.0001
Risk Group	0.68(0.16)	1.98	<0.0001
Race	1.19(0.17)	3.30	<0.0001
Naïve Analysis[b]			
Variable	$\hat{\beta}$(SE)	Estimated OR (95% CI)	P-value
HIV Status	-0.29(0.15)	0.75	0.06
Age	-0.07(0.01)	0.93	<0.0001
Risk Group	1.13(0.15)	3.08	<0.0001
Race	1.18(0.15)	3.25	<0.0001

a. ML based on eqn. (2.23).

b. ML based on eqn. (2.24).

Table 2.11 Change in BV prevalence for women from HERS visits 1 through visit 4 with covariates adjusted (Correction Analysis with Differential and Nondifferential Assumptions Assuming Independent Misclassification).

Main+Internal, Differential*			
Variable	$\hat{\beta}$(SE)	Estimated OR (95% CI)	P-value
HIV Status	0.27(0.31)	1.32	0.38
Age	-0.08(0.02)	0.93	0.0005
Risk Group	0.67(0.31)	1.95	0.03
Race	1.17(0.31)	3.22	0.0011
Main+internal, non-differential†			
Variable	$\hat{\beta}$(SE)	Estimated OR (95% CI)	P-value
HIV Status	-0.31(0.22)	0.73	0.15
Age	-0.10(0.02)	0.90	<0.0001
Risk Group	1.50(0.22)	4.48	<0.0001
Race	1.66(0.23)	5.24	<0.0001

* SE and SP assumed to vary with dichotomized age, HIV risk group, HIV status, race and index for time point (Visit 1 as the reference level) via model (2.25). n_m=466, n_{v1}=n_{v2}=n_{v3}=n_{v4}=60. $\widehat{\sigma_u}^2 = 3.36$.

† No covariates affecting SE and SP; this assumption is not supported by the data (P<0.0001). $\widehat{\sigma_u}^2 = 3.09$.

Table 2.12 Change in BV prevalence for women from HERS visits 1 through visit 4 with covariates adjusted (Correction Analysis with Differential and Nondifferential Assumptions Allowing for Correlated Misclassification).

Main+Internal, Differential, Correlated Misclassification*			
Variable	$\hat{\beta}$(SE)	Estimated OR (95% CI)	P-value
HIV Status	0.35(0.34)	1.32	0.31
Age	-0.08(0.02)	0.93	0.0005
Risk Group	0.67(0.31)	1.95	0.04
Race	1.17(0.31)	3.22	0.01
Main+Internal, Non-Differential, Correlated Misclassification*			
Variable	$\hat{\beta}$(SE)	Estimated OR (95% CI)	P-value
HIV Status	-0.17(0.24)	0.85	0.50
Age	-0.09(0.02)	0.91	<0.0001
Risk Group	1.45(0.25)	4.28	<0.0001
Race	1.45(0.26)	4.25	<0.0001

* SE and SP assumed to vary with dichotomized age, HIV risk group, HIV status, race and index for time point (Visit 1 as the reference level) via model (2.26). n_m=466, n_{v1}=n_{v2}=n_{v3}=n_{v4}=60. $\widehat{\sigma_u}^2 = 2.07$. $\widehat{\sigma_{u^*}}^2 = 2.22$.

† No covariates affecting SE and SP; this assumption is not supported by the data (P<0.0001). . $\widehat{\sigma_u}^2 = 1.95$. $\widehat{\sigma_{u^*}}^2 = 2.66$.

2.4. Discussion

The problem of misclassification has been studied extensively. However, less emphasis has been placed on response misclassification, as compared to predictor misclassification. In particular, for the case with longitudinally collected binary responses that are subject to error, few references have explicitly provided detailed instruction. Our work here primarily builds upon previous work by Neuhaus (17) and Lyles et al (18), while the likelihood derivation relies heavily on general materials in Carroll et al (3).

Our work differs from previous work in several ways. First, we explicitly provide the form of the likelihood for repeatedly measured misclassified responses when there is internal validation data available. Second, we do not restrict the misclassification probabilities to be non-differential. Third, we also provide an accessible computational method to optimize the likelihood. Fourth, we also note that the usual assumption of conditional independence in misclassification can be relaxed if the user does feel the need to take it into account. Previous work by Neuhaus (17) gave a similar likelihood form and pointed out the possibility of optimizing it when misclassification probabilities are determined by a function. However, the work focused more on the non-differential case with either sensitivity analysis or main study only analysis, and no information about incorporating internal validation was given. Lyles et al (18) demonstrated a way to incorporate internal validation for a matched-pair 2×2 table setting with a likelihood-based approach when the outcome is misclassified, but no covariate adjustment was made. Carroll et al (3) is a general textbook with comprehensive information on likelihood specification and validation designs under the topic of mismeasurement, but specific

motivating examples and details, especially details for computation in the context conveyed here, are not provided.

We have shown in detail how to use our approach to incorporate validation data to correct for longitudinal response misclassification. Although our method can incorporate both external and internal validation, we recommend the internal validation design when possible to avoid the unverifiable transportability assumption, gain more precision in estimation, and allow for more flexibility in modeling misclassification probabilities. Both simulation studies and the HERS example indicate that internal validation design is far more favorable than the external validation design when differential misclassification is present.

Neuhaus (17) suggested that when misclassification probabilities depend on covariates via a function, the "closure property" will not hold because the probability of response depends on the covariate through the GLMM model and also through the misclassification probability function. He also stated that in principle the likelihood can be constructed and optimized, but identifiability issues may arise. We have demonstrated that by introducing a second model for SE/SP, the likelihood with internal validation data incorporated can be derived and maximized using commercial software. Both simulation studies and the HERS example suggest that the approach is generally stable and reliable, although in some unlikely cases there may be numerical difficulty.

In Section 2.3.4, we briefly examined the impact of different designs regarding the proportions of various types of internal validation samples on the precision of estimation. Lyles et al (18) provided some general guidance to achieve maximized estimation

efficiency with a fixed total cost for a 2×2 matched-pair table setting. A similar comprehensive cost-efficiency consideration could be an interesting future research topic and will be beneficial in practice.

We have also examined the case when accounting for correlations in misclassification across time points, by simulation studies and by illustrating it in the HERS example. Our studies suggest that the validity in estimating regression coefficients in the primary model is preserved when simply assuming independence in misclassification, even if correlations do exist, but that efficiency may be a trade-off of doing this. Considering much less complicated computation when assuming independence, we leave it for the users to decide whether the correlations should be taken into account or not.

The approach presented here 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. When that is the case, the use of replicates or multiple imperfect diagnostic tools can be helpful to develop a valid correction method. Future work also includes developing semi-parametric alternatives to the parametric approach presented here, with emphasis on accommodating (via study design and analysis) potentially complex differential misclassification procedures.

Chapter 3 Regression Analysis for Differentially Misclassified Binary Covariates

3.1 Univariate Case

3.1.1 Model Specification

We start with the setting of the ordinary generalized linear model. Assume that our primary interest is to characterize a true underlying model as follows:

$$g\{\Pr(Y_i = 1|X_i, \boldsymbol{C_i})\} = \beta_0 + \beta_1 x_i + \boldsymbol{c_i \gamma} \qquad (3.1)$$

in which g is an arbitrary link, X_i is a binary predictor of interest and subject to misclassification, $\mathbf{C_i} = (C_{i1},\ldots,C_{ip})$ is a covariate vector with p dimensions measured without error. $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_p)^T$ is a parameter vector with p dimensions. With X_i misclassified, instead of observing X_i directly, we observe an error-prone binary predictor Z_i instead. Z_i associates with X_i via misclassification probabilities.

When misclassification is non-differential, the key misclassification properties, known as sensitivity (SE) and specificity (SP), are defined as follows:

$$SE = \Pr(Z_i = 1|X_i = 1) \ \ and \ \ SP = \Pr(Z_i = 0|X_i = 0) \qquad (3.2)$$

Here, both SE and SP are constants and independent of other information. For example, $Pr(Z_i = z_i|X_i = x_i, Y_i, \mathbf{C_i}) == \Pr(Z = z|X = x)$.

If misclassification is differential, SE and SP may vary across subjects. Thus, we define them as follows:

$$SE_{y c_i^*} = Pr(Z_i = 1 | X_i = 1, Y_i, \boldsymbol{C_i}^*) \;\; and \;\; SP_{y c_i^*}$$
$$= Pr(Z_i = 0 | X_i = 0, Y_i, \boldsymbol{C_i}^*) \qquad (3.3)$$

where $\mathbf{C_i}^* = (C^*_{i1}, \ldots, C^*_{iq})^T$ is a low-dimensional and quantifiable covariate vector that has impact on SE and SP. $\mathbf{C_i}^*$ may or may not overlap with $\mathbf{C_i}$ in eqn. (3.1). We further assume here misclassification rates only depend on information for the ith subject. More specifically, $Pr(Z_i = 1 | \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{C}^*) = Pr(Z_i = 1 | X_i, \boldsymbol{Y_i}, \boldsymbol{C_i^*})$.

3.1.2 External Validation: Non-differential Misclassification

As defined in 2.1.2, an external validation sample is a sample independent of the main study, and in external validation samples, (X, Z) pairs of (X_i, Z_i) are typically measured. An example of external validation data could be a similar study to one's own that has previously been published in the literature. The nature of external validation sampling requires the assumption of "transportability", i.e., the misclassification properties in the validation sample are the same as those operating in the main study sample (3). In addition, in external validation sampling, it is seldom the case that covariates are also measured. Thus, the use of the external validation design usually forces the assumption of non-differentiality. Advantages of external validation sampling, however, include cost-efficiency and convenience.

Let n_m be the number of subjects in the main study and n_v be the number in the external validation sample. Here we consider modeling the joint probability Pr(Y, Z | **C**, **C****) in eqn. (3.4) instead of Pr(Y|Z, **C, C****):

$$\Pr(Y_i = y_i, Z_i = z_i | \boldsymbol{C_i} = \boldsymbol{c_i}, \boldsymbol{C_i}^* = \boldsymbol{c_i}^*, \boldsymbol{C_i}^{**} = \boldsymbol{c_i}^{**})$$

$$= \sum_{x_i=0}^{1} \Pr(Z_i = z_i | x_i, y_i, \boldsymbol{c_i^*}) \Pr(Y_i = y_i | x_i, \boldsymbol{c_i}) \Pr(X_i = x_i | \boldsymbol{c_i}^{**}) \qquad (3.4)$$

The first term in eqn. (3.4) reflects the SE/SP property, and in the case of non-differential misclassification, this term can be further reduced to $\Pr(Z_i=z_i | X_i=x_i)$ as in eqn. (3.2). The second term in eqn. (3.4) represents the main model in eqn. (3.1). To utilize eqn. (3.4), we introduce an "X|C" model as follows:

$$g^{**}\{\Pr(X_i = 1 | \boldsymbol{C_i})\} = \varphi_0 + \boldsymbol{c_i}^{**}\boldsymbol{\varphi} \qquad (3.5)$$

where g** is an arbitrary link, $\boldsymbol{\varphi}=(\varphi_1,\ldots \varphi_r)$ is parameter vector and $\mathbf{C_i}^{**}$ is an 1×n row vector of covariates that may be a subset of $\mathbf{C_i}$ and or may include other variables. Let $P_{yik}=\Pr(Y_i=1|X_i=k, \mathbf{C_i}=\mathbf{c_i})$ and $P_{xi}=\Pr(X_i=1|\ \mathbf{C_i}^{**}=\mathbf{c_i}^{**})$ where k=(0,1). By integrating all the information together and assuming non-differentiality, we have the likelihood contribution from the main study as follows:

$$L_m = \prod_{i=1}^{n_m} \{[SE \times P_{yi1} \times P_{xi} + (1-SP) \times P_{yi0} \times (1-P_{xi})]^{y_i z_i}$$

$$\times [(1-SE) \times P_{yi1} \times P_{xi} + SP \times P_{yi0} \times (1-P_{xi})]^{y_i(1-z_i)}$$

$$\times [SE \times (1-P_{yi1}) \times P_{xi} + (1-SP) \times (1-P_{yi0}) \times (1-P_{xi})]^{(1-y_i)z_i}$$

$$\times [(1-SE) \times (1-P_{yi1}) \times P_{xi} + SP \times (1-P_{yi0}) \times (1-P_{xi})]^{(1-y_i)(1-z_i)}\} \qquad (3.6)$$

For the external validation sample, we assume that each pair (X_i, Z_i) (i=1, … , n_v) contributes the likelihood in eqn. (3.7):

$$\Pr(Z_i = z_i, X_i = x_i) = \Pr(Z_i = z_i | X_i = x_i)\Pr(X_i = x_i) \qquad (3.7)$$

where the first term reflects the misclassification probability defined by SE/SP, and the second term reflects a nuisance parameter characterizing the prevalence of X in the external validation sampling population. Denote P_{xv}=Pr(X=1) in the validation study population. Thus, the likelihood contribution from the external validation sample is:

$L_v = \prod_{i=1}^{n_v}\{[SE \times P_{xv}]^{x_i z_i} \times [(1-SE) \times P_{xv}]^{x_i(1-z_i)}[(1-SP) \times (1-P_{xv})]^{(1-x_i)z_i} \times [SP \times (1-P_{xv})]^{(1-x_i)(1-z_i)}\}$ (3.8).

The full likelihood is proportional to L_mxL_v.

3.1.3 Internal Validation: Differential Misclassification

Unlike external validation data, an internal validation sample consists of a proportion of subjects randomly selected from the convenient study sample. In this subset, in addition to measurements on Y, Z and other covariates, the true predictor X is also measured. The nature of an internal validation design ensures several benefits, including avoidance of making the assumption of transportability, improved statistical efficiency and the possibility of modeling SE/SP differentially. However, it is clearly more labor-intensive than is using an external validation sample.

Taking similar steps as in Section 3.1.2, we can derive the likelihood contribution for the main study as follows:

$$L_m = \prod_{i=1}^{n_m}\{[SE_{yc_i^*} \times P_{yi1} \times P_{xi} + (1-SP_{yc_i^*}) \times P_{yi0} \times (1-P_{xi})]^{y_i z_i}$$

$$\times [(1-SE_{yc_i^*}) \times P_{yi1} \times P_{xi} + SP_{yc_i^*} \times P_{yi0} \times (1-P_{xi})]^{y_i(1-z_i)}$$

$$\times [SE_{yc_i^*} \times (1-P_{yi1}) \times P_{xi} + (1-SP_{yc_i^*}) \times (1-P_{yi0}) \times (1-P_{xi})]^{(1-y_i)z_i}$$

$$\times [(1 - SE_{y\boldsymbol{c}_i^*}) \times (1 - P_{yi1}) \times P_{xi} + SP_{y\boldsymbol{c}_i^*} \times (1 - P_{yi0}) \times (1 - P_{xi})]^{(1-y_i)(1-z_i)}\} \quad (3.9).$$

Although eqn. (3.9) has a similar form as eqn. (3.6), attention should be paid to the misclassification probability terms. In the case of differential misclassification, SE and SP depend on subject-specific information, indicated by the subscripts. Similarly, we have the likelihood contribution from the validation sample as in eqn. (3.10).

$$L_v = \prod_{i=1}^{n_v} \{[SE_{y\boldsymbol{c}_i^*} \times P_{yi1}{}^{y_i} \times (1 - P_{yi1})^{1-y_i} \times P_{xi}]^{x_i z_i}$$

$$[(1 - SE_{y\boldsymbol{c}_i^*}) \times P_{yi1}{}^{y_i} \times (1 - P_{yi1})^{1-y_i} \times P_{xi}]^{x_i(1-z_i)}$$

$$[(1 - SP_{y\boldsymbol{c}_i^*}) \times P_{yi0}{}^{y_i} \times (1 - P_{yi0})^{1-y_i} \times (1 - P_{xi})]^{(1-x_i)z_i}$$

$$[SP_{y\boldsymbol{c}_i^*} \times P_{yi0}{}^{y_i} \times (1 - P_{yi0})^{1-y_i} \times (1 - P_{xi})]^{(1-x_i)(1-z_i)}\} \quad (3.10).$$

The full likelihood is proportional to $L_m x L_v$. To set the stage for the proposed approach when misclassification is differential, we introduce a third generalized linear model that characterizes the association of SE/SP with covariates Y and $\mathbf{C}^*$:

$$g^*\{\Pr(Z_i = 1|Y_i, X_i, \boldsymbol{C_i}^*)\} = \delta_0 + \sum_{q=1}^{q} \delta_k c_{ik}^* + \delta_{q+1} x_i + \delta_{q+2} y_i \quad (3.11),$$

where g* is an arbitrary link and the predictor in $\mathbf{C_i}^*$ may or may not overlap with those in $\mathbf{C_i}$. Eqn. (3.11) allows flexibility in modeling SE/SP, depending on subject-specific covariate information. Assuming g* is the logistic link, it implies that SE_{yci*} and SP_{yci*} are functions of parameters in eqn. (3.11), as follows:

$$SE_{y\boldsymbol{c}_i^*} = \Pr(\mathrm{Z_i} = 1|\mathrm{X_i} = 1, Y_i, \mathbf{C_i}^*) = \frac{\exp(\eta_{i1})}{1 + \exp(\eta_{i1})}$$

and

$$SP_{yc_i^*} = \Pr(\mathrm{Z_i} = 0|\mathrm{X_i} = 0, Y_i, \mathbf{C_i}^*) = \frac{1}{1+\exp(\eta_{i0})} \qquad (3.12)$$

with $\eta_{\mathrm{iy}} = \delta_0 + \sum_{k=1}^{q} \delta_k c_{ik}^* + \delta_{q+1} x_i + \delta_{q+2} y_i \ \ (\mathrm{y_i} = 0, 1)$.

In practice, a hypothesis test can be conducted to test whether the misclassification is differential or not. Based on main/internal validation study design, if H_0 is not rejected, the likelihood could be reduced to the version that applies when misclassification is non-differential.

3.1.4 Note on Impact of Mis-specifying X|C Model

Model selection can be used to select the covariate vector $\mathbf{C_i}^{**}$ in the X|C model in practice without much technical difficulty. When the X|C model specified in the likelihood differs from the true underlying model, estimates for (β, γ) in the main model (3.1) may be invalid. The validity only preserves when the X|C model is correctly specified or when a covariate omitted from the X|C model is not important for the main model (3.1). Therefore, a careful preliminary model selection on X|C model is recommended to ensure the validity of ML estimates for the main model (3.1) parameters.

3.2 Extension to Repeated Measures

3.2.1 Model Specification

In this section, we move further to consider the situation when the response and the error-prone predictor are measured repeatedly within each subject, while other predictors $\mathbf{C_{ij}}$ may be repeatedly measured or fixed across all occasions. To demonstrate how the

approach proposed can be extended to the case with repeated measures, we consider a generalized linear mixed model:

$$g\{\Pr(Y_{ij}=1|X_{ij},\boldsymbol{C_{ij}},u_{iy})\}=\beta_0+\beta_1 x_{ij}+\boldsymbol{c_{ij}\gamma}+u_{iy} \qquad (3.13)$$

This model is a natural extension of model (3.1) to accommodate repeated measures. In model (3.13), Y_{ij} (0/1)and X_{ij} (0/1) are the repeated outcome measured without error and the predictor variable, both of which would ideally be measured repeatedly at the j-th occasion on the i-th subject (i=1,…,n; j=1,…, J_i). $\mathbf{C_{ij}}$ is a 1×q covariate vector, possibly consisting of a mix of occasion-stationary and occasion-varying predictors. We assume that the random effects u_{iy} represent i.i.d. draws from f(u_{iy}); typically, we assume $u_{iy}\widetilde{i.i.d.}N(0,\ \sigma_{uy}^2)$.

Similar as in the univariate case in Section 3.1, we assume that in the main study that error-rpone Z_{ij} replace X_{ij} (i=1,…, n; j=1,…, J_i). As before, we introduce an X|C model for later likelihood derivation purposes.

$$g^{**}\{\Pr(X_{ij}=1|\boldsymbol{C_{ij}},u_{\ ix}\ )\}=\varphi_0+\boldsymbol{c_{ij}}^{**}\boldsymbol{\varphi}+u_{\ ix} \qquad (3.14)$$

In (3.14) the random effects u_{ix} are assumed to follow f(u_{ix}) ; typically, we assume $u_{ix}\widetilde{i.i.d.}N(0,\ \sigma_{ux}^2)$.

Misclassification probabilities SE and SP are defined as follows when misclassification is non-differential:

$$SE\ =\Pr(Z_{ij}=1|X_{ij}=1)\ \ and\ \ SP\ =\Pr(Z_{ij}=0|X_{ij}=0) \qquad (3.15)$$

With differential misclassification, we allow subject-specific information to have an impact on SE/SP in a fashion similar to that seen in Section 3.1:

$$SE_{y c_{ij}^*} = \Pr\left(\mathrm{Z_{ij}} = 1 \middle| \mathrm{X_{ij}} = 1, \quad \mathrm{y_{ij}}, \mathbf{C_{ij}}^* \right) \; and$$

$$SP_{y c_{ij}^*} = \Pr\left(\mathrm{Z_{ij}} = 0 \middle| \mathrm{X_{ij}} = 0, \mathrm{y_{ij}}, \; \mathbf{C_{ij}}^* \right) \qquad (3.16),$$

and note here that SE and SP may also vary across occasions.

3.2.2 External Validation: Non-Differential Misclassification

As with the univariate case, we start with non-differential misclassification. Following the strategy in section 3.1.2, the likelihood contribution from the main study is derived as follows:

$$L_m = \prod_{i=1}^{n_m} \iint \prod_{j=1}^{J_i} \{[SE \quad \times P_{yij1} \times P_{xij} + (1 - SP \quad) \times P_{yij0} \times (1 - P_{xij})]^{y_{ij} z_{ij}}$$

$$\times [(1 - SE \quad) \times P_{yij1} \times P_{xij} + SP \quad \times P_{yij0} \times (1 - P_{xij})]^{y_{ij}(1 - z_{ij})}$$

$$\times [SE \quad \times (1 - P_{yij1}) \times P_{xij} + (1 - SP \quad) \times (1 - P_{yij0}) \times (1 - P_{xij})]^{(1 - y_{ij}) z_{ij}}$$

$$\times [(1 - SE \quad) \times (1 - P_{yij1}) \times P_{xij} + SP \quad \times (1 - P_{yij0}) \times (1 - P_{xij})]^{(1 - y_{ij})(1 - z_{ij})} \} f(u_{ix}, u_{iy}) du_{ix} du_{iy}$$

where P_{yijk}=Pr(Y_{ij}=1|X_{ij}=k,$\mathbf{C_{ij}}$=$\mathbf{c_{ij}}$) and P_{xij}=Pr(X_{ij}=1| $\mathbf{C_{ij}}^{**}$=$\mathbf{c_{ij}}^{**}$) (k=0, 1). Sometimes X may be a baseline predictor whose value carries across all occasions. When that is the case, the random effects u_{ix} may be dropped from model (3.14) accordingly. We assume that in the external validation sample, only pairs of (Z_i, X_i) (i=1 ,…, n_v)are observed. Then the validation data contributes the same likelihood as in eqn. (3.8) because the layout of the external validation sample is the same. Then as before, the full likelihood is

proportional to $L_m x L_v$. Note here we do not put any constraints on the correlation of u_{iy} and u_{ix}, so a covariance term σ_{xy} can be introduced and modeled if needed.

3.2.3 Internal Validation: Differential

In order to model the dependency of SE/SP on covariate information, a model in eqn. (3.18) could be assumed.

$$g^*\{\Pr(Z_{ij}=1|Y_{ij},X_{ij},\boldsymbol{C_{ij}}^*,u_{iz})\} = \delta_0+\sum_{k=1}^{q}\delta_k c_{ijk}^* + \delta_{q+1}x_{ij}+\delta_{q+2}y_{ij}+u_{iz} \quad (3.18)$$

Note that this is analogous to model (3.11), except with the random effects uiz to account for repeated sampling of (Z_{ij}, X_{ij}) in the validation set. Typically, we assume $u_{iz}\widetilde{i.i.d.}N(0,\sigma_{uz}^2)$. In practical setting, assuming independent misclassification processes across time is often sensible, reflected in eqn.(3.19).

$$g^*\{\Pr(Z_{ij}=1|Y_{ij},X_{ij},\boldsymbol{C_{ij}}^*)\} = \delta_0+\sum_{k=1}^{q}\delta_k c_{ijk}^* + \delta_{q+1}x_{ij}+\delta_{q+2}y_{ij} \quad (3.19)$$

Thus, from now on, we will make this assumption unless specified otherwise. $SE_{yc_{ij}^*}$ and $SP_{yc_{ij}^*}$ can be defined accordingly based on eqn.(3.19). When misclassification is differential, the main study contributes to the likelihood via eqn. (3.20).

$$L_m=\prod_{i=1}^{n_m}\iint\prod_{j=1}^{n_j}\{[SE_{yc_{ij}^*}\times P_{yij1}\times P_{xij}+\left(1-SP_{yc_{ij}^*}\right)\times P_{yij0}\times(1-P_{xij})]^{y_{ij}z_{ij}}$$

$$\times[(1-SE_{yc_{ij}^*})\times P_{yij1}\times P_{xij}+SP_{yc_{ij}^*}\times P_{yij0}\times\left(1-P_{xij}\right)]^{y_{ij}(1-z_{ij})}$$

$$\times [SE_{yc_{ij}^*} \times (1 - P_{yij1}) \times P_{xij} + (1 - SP_{yc_{ij}^*}) \times (1 - P_{yij0}) \times (1 - P_{xij})]^{(1-y_{ij})z_{ij}}$$

$$\times \left[\left(1 - SE_{yc_{ij}^*}\right) \times (1 - P_{yij1}) \times P_{xij} + SP_{yc_{ij}^*} \times (1 - P_{yij0}) \times (1 - P_{xij})\right]^{(1-y_{ij})(1-z_{ij})}\Big\}$$

$$f(u_x, u_y)du_x du_y \qquad (3.20).$$

The likelihood for the internal validation data is

$$L_v = \prod_{i=1}^{n_v} \int \int \prod_{j=1}^{n_j} \{[SE_{yc_{ij}^*} \times {P_{yij1}}^{y_{ij}} \times (1 - P_{yij1})^{1-y_{ij}} \times P_{xij}]^{x_{ij}z_{ij}}$$

$$\times [(1 - SE_{yc_{ij}^*}) \times {P_{yij1}}^{y_{ij}} \times (1 - P_{yij1})^{1-y_{ij}} \times P_{xij}]^{x_{ij}(1-z_{ij})}$$

$$\times [(1 - SP_{yc_{ij}^*}) \times {P_{yij0}}^{y_{ij}} \times (1 - P_{yij0})^{1-y_{ij}} \times (1 - P_{xij})]^{(1-x_{ij})z_{ij}}$$

$$\times [SP_{yc_{ij}^*} \times {P_{yij0}}^{y_{ij}} \times (1 - P_{yij0})^{1-y_{ij}} \times (1 - P_{xij})]^{(1-x_{ij})(1-z_{ij})}\}$$

$$f(u_x, u_y)du_x du_y \qquad (3.21).$$

Thus, the full likelihood is proportional to the product of L_m and L_v.

3.2.4 Estimation

The marginal full likelihood can be optimized after integrating out the random effects. Numerical methods available for the integration computation include adaptive Gaussian quardrature (46) and the first-order method (47). The full likelihood can be then optimized via quasi-Newton optimization, and standard errors of estimates may be obtained from the final Hessian matrix. Such optimization techniques are well developed in standard software as optimization routines. All simulations and data examples are carried out using the NLMIXED procedure in SAS 9.2 (48).

3.3. Simulation Studies

In this section, we describe simulation studies to assess the performance of the proposed ML approach under different situations. In all cases we confine our attention to the cases where g, g* and g** are logit links. It can be easily generalized to other links without much difficulty.

3.3.1 External Validation in Univariate Case: Non-Differential Misclassification

Table 3.1 summarizes simulations under external validation sampling with a univariate predictor nondifferentially misclassified. In each case considered, data were generated via the logistic version of model (3.1) containing a binary predictor (X) that is misclassified as Z in the observed main study data, a continuous covariate (C_1) and a binary covariate (C_2). The version of the X|C model (eqn. (3.5)) included both C_1 and C_2 as predictors. Three models were examined, each with $(\beta_0, \beta_1, \gamma_1, \gamma_2, \varphi_0, \varphi_1, \varphi_2)$=(0 , 1, 1.5, -1, 0, -0.25, 0.5) and sample size of (n_m, n_v)=(1000,200). The true SE and SP were 0.7 and 0.85 respectively. C_1 was normally distributed with mean 0 and variance 4. C_2 was a Bernoulli variable with a probability of 0.5. A total of 500 simulations were run.

Unsurprisingly, the "naïve" analysis produced drastically attenuated estimates for all parameters in model (3.1). In general, the results show that the ML method proposed in Section 3.1 performs well. As expected, the efficiency of β_1 suffers as compared to the ideal analysis, while γ_1 and γ_2 are estimated with only slightly lower precision than in the ideal analysis.

Table 3.1 Results of simulations comparing ML estimates under external validation sampling for logistic regression with non-differential predictor misclassification*

Main + External Validation				
Parameter	Mean	SD	95% CI Coverage	Convergence
β_0	-0.04	0.30	0.96	500
β_1	1.06	0.48	0.96	500
γ_1	1.53	0.13	0.97	500
γ_2	-1.01	0.23	0.95	500
φ_0	0.02	0.29	0.97	500
φ_1	-0.26	0.08	0.95	500
φ_2	0.52	0.28	0.96	500
Ideal Analysis				
Parameter	Mean	SD	95% CI Coverage	Convergence
β_0	-0.01	0.17	0.93	500
β_1	1.02	0.20	0.96	500
γ_1	1.52	0.10	0.97	500
γ_2	-1.00	0.21	0.94	500
Naïve Analysis				
Parameter	Mean	SD	95% CI Coverage	Convergence
β_0	0.27	0.16	0.56	500
β_1	0.51	0.20	0.29	500
γ_1	1.43	0.09	0.84	500
γ_2	-0.89	0.20	0.87	500

* 500 simulations; $(\beta_0, \beta_1, \gamma_1, \gamma_2, \varphi_0, \varphi_1, \varphi_2)=(0, 1, 1.5, -1, 0, -0.25, 0.5)$; SE=0.7, SP=0.85; $n_m=1000$, $n_v=200$; $C_1 \sim N(0,4)$; $C_2 \sim Ber(0.5)$.

3.3.2 Internal Validation in Univariate Case: Differential Misclassification

Table 3.2 summarizes simulation results under internal validation sampling with differential misclassification in the univariate case. Data were generated via the logistic versions of model (3.1) and model (3.5) as described in section 3.3.1. To allow for differential misclassification, a logistic version of model (3.11) was specified as the SE/SP model with X, Y and C_2 as predictors, and $(\delta_0, \delta_1, \delta_2, \delta 3)=(-1.5, 0.2, -0.1, 2.6)$. As in Section 3.1, 1000 main study observations with 200 validation sample observation were used, and three models were examined. A total of 500 simulations were run.

Again, the "naïve" analysis produces largely attenuated estimates for parameters of interest. The ML method proposed with internal validation data incorporated performs well with regard to point estimates of the parameters of interest ($\beta_0, \beta_1, \gamma_1, \gamma_2$). There is some efficiency loss in the estimation as expected. The parameters characterizing the SE/SP model and the X|C model are also estimated reliably.

Table 3.2 Results of simulations comparing ML estimates under internal validation sampling for univariate logistic regression with differential predictor misclassification*

Main + Internal Validation				
Parameter	Mean	SD	95% CI Coverage	Convergence
β_0	-0.08	0.42	0.94	500
β_1	1.17	0.77	0.92	500
γ_1	1.57	0.17	0.96	500
γ_2	-1.04	0.30	0.96	500
φ_0	0.02	0.60	0.96	500
φ_1	-0.27	1.05	0.96	500
φ_2	0.52	0.29	0.94	500
δ_0	-1.51	0.36	0.93	500
δ_1	0.22	0.37	0.91	500
δ_2	2.62	0.38	0.95	500
δ_3	-0.10	0.33	0.94	500
True Predictor				
Parameter	Mean	SD	95% CI Coverage	Convergence
β_0	0.00	0.16	0.94	500
β_1	1.01	0.19	0.94	500
γ_1	1.52	0.09	0.95	500
γ_2	-1.01	0.18	0.95	500
Naive				
Parameter	Mean	SD	95% CI Coverage	Convergence
β_0	0.28	0.15	0.50	500
β_1	0.47	0.18	0.16	500
γ_1	1.44	0.08	0.86	500
γ_2	-0.92	0.17	0.92	500

* 500 simulations; $(\beta_0, \beta_1, \gamma_1, \gamma_2, \varphi_0, \varphi_1, \varphi_2)$=(0 , 1, 1.5, -1, 0, -0.25, 0.5); n_m=1000, n_v=200;C_1~N(0,4);C_2~Ber(0.5). $(\delta_0, \delta_1, \delta_2, \delta_3)$=(-1.5, 0.2, 2.6, -0.1).

3.3.3 External Validation in Longitudinal Case: Non-Differential Misclassification

Table 3.3 presents the results of simulations designed to assess the implementation and performance of ML in the setting of a repeated measures study based on the methods in Section 3.2. For illustrative purposes, we considered the case where there were two repeated measurements. Data were simulated under the following versions of models (3.21) and (3.22), respectively:

$$logit\{\Pr(Y_{ij}=1|X_i,\boldsymbol{C_i},u_{iy})\}=\beta_0+\beta_1 x_i+\gamma_1 c_{i1}+\gamma_2 t_{ij}+u_{iy} \quad (3.21)$$

(t=0, 1 as the index for time) and

$$logit\{\Pr(X_i=1|\boldsymbol{C_i})\}=\varphi_0+\varphi_1 c_{i1}+\varphi_2 c_{i2} \quad (3.22).$$

The overall sample size for each simulated dataset was 1000 for the main study and 200 for the external validation set. Here C_1 was normally distributed with mean 0 and variance 4, and C_2 followed a Bernoulli distribution of probability 0.5. X was binary and misclassified as Z. C_1, C_2 and X were time-invariant. (β_0, β_1, γ_1, γ_2, φ_0, φ_1, φ_2)=(0 , 1, 1.5, -1, 0, -0.25, 0.5) and the true SE and SP were 0.7 and 0.85 respectively. Five scenarios, the ideal analysis, the naïve analysis, the analysis with X|C correctly specified, the analysis with C_1 left out from the X|C model and the analysis with C_2 left out, were examined.

Table 3.3 Results comparing ML estimates under external validation sampling for pairwise correlated measurements with non-differentially misclassified predictor X and assessing effects of omitted predictor in X | C model *

Main+Internal (X|C correctly specified)			
Parameter	Mean(SD)	Mean StdErr	95% Coverage
β_0	0.01(0.20)	0.19	0.94
β_1	0.98(0.30)	0.29	0.95
γ_1	1.49(0.12)	0.12	0.92
γ_2	-0.99(0.14)	0.14	0.96
Main+Internal (X|C incorrectly specified with C_1 left out)			
Parameter	Mean(SD)	Mean StdErr	95% Coverage
β_0	0.22(0.26)	0.23	0.74
β_1	0.62(0.39)	0.35	0.70
γ_1	1.29(0.09)	0.09	0.34
γ_2	-1.00(0.14)	0.14	0.96
Main+Internal (X|C incorrectly specified with C_2 left out)			
Parameter	Mean(SD)	Mean StdErr	95% Coverage
β_0	0.01(0.22)	0.21	0.94
β_1	0.98(0.34)	0.33	0.94
γ_1	1.49(0.12)	0.13	0.94
γ_2	-0.99(0.14)	0.14	0.96
True predictor			
Parameter	Mean(SD)	Mean StdErr	95% Coverage
β_0	-0.01(0.16)	0.16	0.96
β_1	0.99(0.22)	0.23	0.96
γ_1	1.49(0.11)	0.11	0.94
γ_2	-0.98(0.13)	0.14	0.96
Naive			
Parameter	Mean(SD)	Mean StdErr	95% Coverage
β_0	0.44(0.13)	0.13	0.07
β_1	0.28(0.16)	0.16	0.01
γ_1	1.27(0.08)	0.08	0.19
γ_2	-0.99(0.14)	0.14	0.20

* 500 simulations; $(\beta_0, \beta_1, \gamma_1, \gamma_2, \varphi_0, \varphi_1, \varphi_2)$=(0 , 1, 1.5, -1, 0, -0.25, 0.5); SE=0.7, SP=0.85; n_m=1000, n_v=200; C_1~N(0,4);C_2~Ber(0.5),t=0/1. $(\delta_0, \delta_1, \delta_2, \delta_3)$=(-1.5, 0.2, -0.1, 2.6).

The results indicate that the ML approach proposed performs well when the X|C model is correctly specified. Slight efficiency loss is observed when using the ML approach as expected. The impact of mis-specifying the X|C model depends on the situation. The covariate C_1 is an important covariate for both model (3.21) and model (3.22). Omitting C_1 from the X|C model causes inconsistency in estimates of both β_1 and γ_1. In contrast, C_2 is only a predictor for model (3.22) and is not involved in the main model (3.21). Omitting C_2 in specifying the X|C model reveals no inconsistency in estimates for parameters of interest.

3.3.4 Internal Validation in Longitudinal Case: Differential Misclassification

We considered the same version of model (3.21) and model (3.22) as in Section 3.3.3 with the same true values. Unlike in the nondifferential case, here we specified the SE/SP model (3.23) to allow misclassification probabilities to vary across subjects.

$$logit\{\Pr(Z_{ij} = 1|Y_{ij}, X_{ij}, \boldsymbol{C_{ij}}^*)\} = \delta_0 + \delta_1 c_{ij2} + \delta_2 x_{ij} + \delta_3 y_{ij} \qquad (3.23)$$

with true values (δ_0, δ_1, δ_3, δ_2)=(-1.5, 0.2, -0.1, 2.6). All covariates in consideration here were time-invariant, except for X. The sample size for the main study was 1000, and 200 subjects were randomly assigned into the internal validation sample. Four scenarios, the ideal analysis, the naïve analysis, the analysis with X|C correctly specified, and the analysis with C_2 left out, were examined.

Table 3.4 Results comparing ML estimates under internal validation sampling for pairwise correlated measurements with differentially misclassified predictor X and assessing effects of omitted predictor in X | C model *

Main+Internal (X|C correctly specified)			
Parameter	Mean(SD)	Mean SD	95% Coverage
β_0	-0.03(0.41)	0.36	0.93
β_1	1.05(0.73)	0.66	0.94
γ_1	1.54(0.16)	0.14	0.96
γ_2	-1.02(0.25)	0.28	0.97
φ_0	0.04(0.39)	0.37	0.97
φ_1	-0.28(0.15)	0.13	0.96
φ_2	0.47(0.56)	0.54	0.96
δ_0	-1.57(0.38)	0.36	0.94
δ_1	0.23(0.35)	0.34	0.94
δ_2	2.66(0.37)	0.37	0.95
δ_3	-0.07(0.36)	0.33	0.93
Main+Internal (X|C incorrectly specified)			
Parameter	Mean(SD)	Mean SD	95% Coverage
β_0	-0.07(0.41)	0.38	0.95
β_1	1.05(0.71)	0.65	0.94
γ_1	1.54(0.16)	0.14	0.96
γ_2	-0.95(0.20)	0.21	0.93
φ_0	0.27(0.29)	0.27	0.85
φ_1	-0.28(0.16)	0.12	0.96
φ_2	-1.68(0.37)	0.35	0.91
δ_0	2.61(0.36)	0.35	0.95
δ_1	-0.07(0.35)	0.32	0.93
δ_2	0.50(0.17)	0.17	0.58
True Predictor			
Parameter	Mean(SD)	Mean SD	95% Coverage
$\beta 0$	0.00(0.15)	0.15	0.97
$\beta 1$	1.00(0.18)	0.18	0.95
$\gamma 1$	1.52(0.09)	0.09	0.96
$\gamma 2$	-1.01(0.17)	0.18	0.96

* 500 simulations; $(\beta_0, \beta_1, \gamma_1, \gamma_2, \varphi_0, \varphi_1, \varphi_2)$=(0 , 1, 1.5, -1, 0, -0.25, 0.5); n_m=1000, n_v=200; C_1~N(0,4);C_2~Ber(0.5),t=0/1. $(\delta_0, \delta_1, \delta_2, \delta 3)$=(-1.5, 0.2, 2.6, -0.1).

Table 3.4 illustrates the performance of the different models. When the X|C model is correctly specified, the proposed ML approach performs quite well with only slight efficiency loss as compared to the ideal analysis. In this example, C_2 is not involved in the main model (3.21), but is an important predictor for the X|C model (3.22) and the SE/SP model (3.23). When C_2 is omitted when specifying the X|C model in the likelihood, maximization leads to invalid estimates for δ_1 in the SE/SP model. However, parameters of primary interest (β_0, β_1, γ_1, γ_2) are still appear to be reliably estimated. Although care should always be taken when selecting the X|C model, the results shown here indicate that the X|C model does have some robustness to mis-specification when it comes to estimating the primary model parameters.

3.4. Example

3.4.1 HERS Example

The HERS example was described in Section 1.4. Here we considered elevated vaginal PH (>4.7) as a response (Y) of interest for illustrative purposes, and the predictor variable richomoniasis status (X) may be misclassified when diagnosed by the wet mount method. The gold standard assessment of trichomoniasis is culture testing, which was also measured along with wet mount for all subjects in HERS.

3.4.2 Example 1: Univariate Analysis with Visit 4

A total of 873 women with all measurements on the response, predictor and covariates available at visit 4 were considered. Among them, 18% of women had trichomonads present in culture testing, while only 7.7% had positive wet mount results. 53% of women were observed to have vaginal PH greater than 4.7 (PH). 61.5% of the women

were blacks, 53.0% were intravenous drug users (IDU) and 67.8% were HIV positive. One-fourth of the women were randomly selected as an internal validation subset to illustrate misclassification correction, while for the remaining three-fourths included in the main study sample, trichomoniasis diagnoses from culture testing were ignored.

With a careful model selection, the chosen version of eqn. (3.1) using the true (CULTURE) diagnosis is:

$$logit[\Pr(PH = 1)] = \beta_0 + \beta_1 RISKGRP + \beta_2 CULTURE \qquad (3.24)$$

We then fit the same model by substituting the error-prone wet mount (WET) as the predictor

$$logit[\Pr(PH = 1)] = \beta_0 + \beta_1 RISKGRP + \beta_2 WET \qquad (3.25).$$

The results of these two analyses are summarized in the upper half of Table 3.5. The results differ markedly in terms of the magnitude of the estimated OR for trichomoniasis (1.54 for eqn. (3.24) and 4.14 for eqn. (3.25)).

Model selection via eqn. (3.11) fit to women in the internal validation subsample revealed that the outcome status has a significant impact on the performance of the wet mount test.

$$logit[\Pr(WET = 1)] = \theta_0 + \theta_1 CULTURE + \theta_2 PH + \theta_3 RISKGRP \qquad (3.26),$$

indicating the presence of differential misclassification. Similarly, fitting eqn. (3.5) to all 916 women yielded a version of eqn. (3.27) used in the likelihood:

$$logit[\Pr(CULTURE = 1)] = \gamma_0 + \gamma_1 RACE + \gamma_2 RISKGRP + \gamma_3 AGE \qquad (3.27)$$

Table 3.5 Results of maximum likelihood analysis of main/internal validation study data on 873 women (n_m=655, n_v=218) at the 4th visit: estimates of primary analyses.

Ideal Analysis[1]			
Variable	$\hat{\beta}$ (std. error)	Estimated OR 95% CI	*P*-value
Risk Group (IDU vs. sex)	0.43 (0.14)	1.42 (0.99, 2.02)	0.05
trichomoniasis (Culture) (yes vs no)	0.35 (0.18)	1.54 (1.18, 2.02)	0.002
Naïve Analysis[2]			
Variable	$\hat{\beta}$ (std. error)	Estimated OR 95% CI	*P*-value
Risk Group (IDU vs. sex)	0.38 (0.14)	1.25 (1.12, 1.93)	0.01
trichomoniasis (Wet Mount) (yes vs no)	1.42 (0.33)	4.14 (2.18, 7.87)	<0.0001
Assuming differential misclassification[3]			
Variable	$\hat{\beta}$ (std. error)	Estimated OR 95% CI	*P*-value
Risk Group (IDU vs. sex)	0.45 (0.14)	1.57 (1.14, 1.99)	0.001
trichomoniasis (yes vs no)	0.33 (0.29)	1.39 (0.60, 2.17)	0.25
Assuming non-differential misclassification[4]			
Variable	$\hat{\beta}$ (std. error)	Estimated OR 95% CI	*P*-value
Risk Group (IDU vs. sex)	0.42 (0.14)	1.52 (1.10,1.94)	0.003
trichomoniasis (yes vs no)	0.88 (0.26)	2.41 (1.17, 3.65)	0.001

1. Analysis using eqn. (3.24). 2. Analysis using eqn. (3.25). 3. Analysis using models (3.25), (3.26) and (3.27). 4. Analysis using models (3.25), (3.26) and (3.28).

The lower half of Table 3.5 summarizes a complete analysis of data via the joint likelihood of eqn.s (3.25)-(3.27) or (3.25), (3.26) and (3.28) when assuming nondifferential misclassification.

$$logit[\Pr(WET = 1)] = \theta_0 + \theta_1 CULTURE \qquad (3.28)$$

To test whether the assumption of nondifferentiality is plausible or not (restricting θ_2=0 in eqn. (3.26)), a likelihood ratio test was performed, with a significant result supporting differential misclassification (χ^2=12.7, p<0.01). This conclusion is also supported by the analytic results. By allowing differential misclassification, the magnitude of the estimated OR of trichomoniasis is closer to that from the ideal analysis. In contrast, if nondifferentiality is inappropriately assumed, the analytic results are similar to those of the naïve analysis, with an elevated estimate for trichomoniasis.

3.4.3 Example 2: Longitudinal Analysis

In the second example, we use data from black, white and Hispanic patients from the 4th and 5th semi-annual visits. In total, 734 women were considered. Similarly as in Section 3.4.2, one-fourth of the women were randomly selected as internal validation subset. We still consider vaginal PH greater than 4.7 as the response variable. After model selection, the chosen version of model (3.13) is:

$$logit[\Pr(PH = 1)] = \beta_0 + \beta_1 RISKGRP + \beta_2 CULTURE + u_{iy} \qquad (3.29)$$

We also fit eqn. (3.29) by replacing culture testing results with wet mount results.

The results of these two analyses are summarized in the upper half of Table 3.6. Similarly as in the univariate analysis presented in Section 3.4.2, the results differ primarily in

terms of the magnitude of the estimated OR for trichomoniasis (2.11 for the gold standard analysis and 8.25 for the naïve analysis).

The selected version of model (3.20) by allowing independent and differential misclassification is,

$$logit[\Pr(WET = 1)] = \theta_0 + \theta_1 CULTURE + \theta_2 PH + \theta_3 RISKGRP \qquad (3.30),$$

where misclassification in diagnosing trichomoniasis is differential about the response variable as well as the HIV risk cohort. Note that we assume the misclassification process at each visit is independent from that at other visits.

Similarly, the selected version of model (3.13) based on evaluating X|C models with the total sample size yielded eqn. (3.31):

$$logit[\Pr(CULTURE = 1)] = \gamma_0 + \gamma_1 RACE + \gamma_2 RISKGRP + u_{ix} \qquad (3.31)$$

The lower half of Table 3.6 summarizes a complete analysis of data via the joint likelihood of eqn.s (3.29)-(3.31). As in section 3.4.2, misclassification is significantly differential (χ^2=11.5, p<0.01). Therefore, we do not further reduce the model and keep it as it is. The differential model produces results closes to those of the ideal analysis. In contrast to the differential model, the nondifferential model yields highly biased estimate for trichomoniasis, although there appears markedly improvement compared to the naïve analysis.

Table 3.6 Results of maximum likelihood analysis of main/internal validation study data on 734 women (n_m=550, n_v=184) at the 4th and 5th visit: estimates of primary analyses.

Ideal Analysis[1]			
Variable	$\hat{\beta}$ (std. error)	Estimated OR 95% CI	*P*-value
Risk Group (IDU vs. sex)	0.28 (0.14)	1.33 (0.97, 1.69)	0.04
trichomoniasis (Culture) (yes vs no)	0.75 (0.20)	2.11 (1.27, 2.96)	0.002
Naïve Analysis[2]			
Variable	$\hat{\beta}$ (std. error)	Estimated OR 95% CI	*P*-value
Risk Group (IDU vs. sex)	0.23 (0.14)	1.26 (0.92, 1.59)	0.09
trichomoniasis (Wet Mount) (yes vs no)	2.11 (0.43)	8.25 (1.29, 15.21)	<0.0001
Assuming differential misclassification[3]			
Variable	$\hat{\beta}$ (std. error)	Estimated OR 95% CI	*P*-value
Risk Group (IDU vs. sex)	0.28 (0.24)	1.32 (0.82, 2.13)	0.22
trichomoniasis (yes vs no)	0.85 (0.69)	2.35 (0.61, 9.06)	0.26
Assuming nondifferential misclassification[3]			
Variable	$\hat{\beta}$ (std. error)	Estimated OR 95% CI	*P*-value
Risk Group (IDU vs. sex)	0.25 (0.14)	1.28 (0.93, 1.63)	0.08
trichomoniasis (yes vs no)	1.42 (0.34)	4.14 (1.37, 6.91)	<0.0001

1. Analysis using culture testing as a predictor in model (3.29). 2. Analysis using wet mount as a predictor as a predictor in model (3.29). 3. Analysis based on internal validation data via models (3.29), (3.30) and (3.31). $\widehat{\sigma^2_{uiy}} = 0.97$, $\widehat{\sigma^2_{uix}} = 0.001$. 4. Analysis based on internal validation data via models (3.29), (3.30) (θ_2=θ_3=0) and (3.31). $\widehat{\sigma^2_{uiy}} = 1.03$, $\widehat{\sigma^2_{uix}} = 0.001$.

3.5. Discussion

In this chapter, we have proposed a parametric method to correct for the bias that stems from predictor misclassification. The primary novelty of the proposed approach lies in a clear illustration of specifying the likelihood functions when internal validation is available. We also provide detailed guidance on how to adjust the likelihood functions when differential misclassification is present. It has also been shown that the approach could be extended to handle a repeatedly measured exposure variable that is subject to misclassification in a generalized linear mixed model context. Although throughout we have used logit links for the response, misclassification and X|C models, other links can be also adopted without too much conceptual and technical difficulty.

We have evaluated the performance of the proposed method via extensive simulations and detailed analysis of trichomoniasis data in the HERS study, both of which have highlighted the value of internal validation sampling, which makes it possible to flexibly model complex differential misclassification. Note that if correction is made based on an inappropriate assumption on the misclassification mechanism, only marginal improvement over the naïve analysis would be offered, as shown in the HERS example, reinforcing the importance of carefully evaluating misclassification mechanisms in practice.

It has been shown that with correctly-specified models, our approach produces reliable estimates on primary parameters of interest. However, it has also been pointed out in Section 3.1.4 that careful model selection is necessary to ensure the validity of primary analyses. Although we recommend careful model selection on the X|C model to ensure

the validity, empirical evidence suggests that the proposed approach maintains validity for estimating primary parameters that are not associated with the X|C model.

We have mainly focused on the case when the misclassification process is independent across occasions in a repeatedly measured study in this chapter. However, it is straightforward to extend the work to accommodate correlated misclassification processes when needed, as shown in eqn. (3.18). Involving more random effects tend to encounter more numerical issues when optimizing the likelihood. Thus, future work could involve development of semiparametric (33) or nonparametric approaches (55) to estimate the misclassification process, which would have great value by making the approach more robust and computationally easy.

Chapter 4 Misclassification in Response and Predictor Variables in 2×2 Tables

4.1 Methods

4.1.1 Notations and Terminology

4.1.1.1 Differential and Dependent Misclassification

Consider a 2 × 2 table in which one intends to measure an error-prone surrogate X^* for a true exposure X and an error-prone surrogate response Y^* for a true response Y. We assume X, X^*, Y and Y^* are all binary variables. Now define $\pi_{ij} = \Pr(X = i, Y = j)$ and $\pi_{ij}^* = \Pr(X^* = i, Y^* = j)$ (i, j=0,1). The true OR of primary interest is calculated as $\pi_{11}\pi_{00}/\pi_{10}\pi_{01}$, while with misclassification in both variables, the naïve OR is $\pi_{11}^*\pi_{00}^*/\pi_{10}^*\pi_{01}^*$.

The observed-data likelihood can be expressed as follows without losing generality:

$$\pi_{ij}^* = \sum_e \sum_d Pr(Y^* = j, X^* = i, Y = d, X = e)$$

$$= \sum_e \sum_d Pr(Y^* = j | Y = d, X = e, X^* = i) Pr\,(X^* = i | X = e, Y = d) \pi_{ed} \quad (4.1)$$

where d, e=0, 1. The first and second terms in eqn. (4.1) represent the most general form of the likelihood expressed with the familiar misclassification parameters known as sensitivity and specificity. With no additional constraints, we define $SE_{yei} = \Pr(Y^* =$

$1|Y=1,X=e,X^*=i)$ and $SP_{yei}=\Pr(Y^*=0|Y=0,X=e,X^*=i)$. Note that misclassification parameters on Y depend on the joint distribution of (X, X^*), which can be viewed as a generalized version of the typical notion of differential misclassification. Similarly, denote $SE_{xd}=\Pr(X^*=1|X=1,Y=d)$ and $SP_{xd}=\Pr(X^*=0|X=0,Y=d)$, taking the typical form of differential misclassification. Eqn. (4. 1) also allows flexible modeling on dependence of misclassification of Y on misclassification of X. With the most general form in eqn. (4.1), note that

$$Pr(Y^*=j,X^*=i|Y=d,X=e)=Pr(Y^*=j|Y=d,X=e)Pr(X^*=i|X=e,Y=d)$$

is not necessarily true, while such an important misclassification assumption is commonly made in previous literature. Thus, for convenience, we consider a misclassification process as in eqn. (4.1) as reflecting "differential and dependent misclassification".

Alternatively, one may also choose to parameterize the observed likelihood in terms of positive and negative predictive values, as reflected here:

$$\pi_{ed}=\sum_i \sum_j Pr(Y^*=j,X^*=i,Y=d,X=e)$$

$$=\sum_i\sum_j Pr(Y=d|Y^*=j,X^*=i,X=e)Pr\,(X=e|X^*=i,Y^*=j)\pi^*_{ij} \quad (4.2)$$

where the first and second terms relate to predictive values of X and Y, defined as $PPV_{yei}=\Pr(Y=1|Y^*=1,X=e,X^*=i)$, $NPV_{yei}=\Pr(Y=0|Y^*=0,X=e,X^*=i)$, $PPV_{xd}=\Pr(X=1|X^*=1,Y^*=d)$ and $NPV_{xd}=\Pr(X=0|X^*=0,Y^*=d)$. In contrast to the parameterization using SE and SP, note that the predictive values of X

depend on the surrogate response measurement. Again, predictive values of Y depend on the joint distribution of (X, X^*), implying the dependence on the other misclassified variable. Note that, when only the exposure X is subject to misclassification, eqn. (4.2) can be rewritten as

$$\pi_{ed} = \sum_i Pr(Y = d, X^* = i, X = e) = \sum_i \Pr(X = e|X^* = i, Y = d)\Pr(X^* = i|Y = d)$$

which conforms to Marshall's proposal (8), from which the term "inverse matrix method" was derived.

4. 1.1.2 Differential and Independent Misclassification

In practice, differential but independent misclassification is of general interest. If taking the parameterization based on SE and SP, this corresponds to reducing eqn. (4.1) to the form as follows:

$$\pi^*_{ij} = \sum_e \sum_d Pr(Y^* = j, X^* = i, Y = d, X = e)$$

$$= \sum_e \sum_d Pr(Y^* = j|Y = d, X = e)Pr\,(X^* = i|X = e, Y = d)\pi_{ed} \quad (4.3)$$

where misclassification on Y only depends on true exposure X characterized by parameters $SE_{ye} = \Pr(Y^* = 1|Y = 1, X = e)$ and $SP_{ye} = \Pr(Y^* = 0|Y = 0.X = e)$. The model for misclassification on X stays the same as in section 4.1.1.1. Assuming independence in misclassification, we imply that $Pr(Y^* = j, X^* = i|Y = d, X = e) = Pr(Y^* = j|Y = d, X = e)Pr(X^* = i|X = e, Y = d)$ holds.

4.*1.1.3 Non-Differential and Independent Misclassification*

When nondifferentiality and independence in misclassification are assumed simultaneously, by defining $SE_x = \Pr(X^* = 1|X = 1)$, $SP_x = \Pr(X^* = 0|X = 0)$, $SE_y = \Pr(Y^* = 1|Y = 1)$ and $SP_y = \Pr(Y^* = 0|Y = 0)$, we can rewrite the observed data likelihood as:

$$\pi_{ij}^* = \sum_e \sum_d Pr(Y^* = j, X^* = i, Y = d, X = e)$$

$$= \sum_e \sum_d Pr(Y^*|Y = d) Pr\,(X^*|X = e)\pi_{ed} \quad (4.4)$$

4.*1.1.4. Other Combinations*

Sections 4.1.1.1-4.1.1.3 give examples of three situations. However, in practice, it may be possible for Y to be differentially but X to be nondifferentially misclassified. Other combinations can exist. For illustrative purposes, we confine our attention to the three situations described above in sections 4.1.1,1-4.1.1.4. The method for other situations can be generalized without conceptual difficulty.

4.1.2 Maximum Likelihood (ML) Approach

In general, the main study likelihood based on observed data paris (Y_i^*, X_i^*) (1=1,…,n_m) can be expressed as:

$$L_m = \prod_{i=1}^{n_m} \pi_{11}^{*(y_i^* x_i^*)} \pi_{01}^{*((1-x_i^*)y_i^*)} \pi_{10}^{*(x_i^*(1-y_i^*))} \pi_{00}^{*((1-x_i^*)(1-y_i^*))} \quad (4.5)$$

where the π^*s take appropriate forms corresponding to different assumptions on the misclassification process described in Section 4.1.1.

For instance, if parameterizing in terms of SE/SP and allowing for differential and dependent misclassification, $\pi_{11}^* = SE_{y11}\pi_{11}SE_{x1} + SE_{y10}\pi_{01}(1-SP_{x1}) + (1-SP_{y11})\pi_{10}SE_{x0} + (1-SP_{y10})\pi_{00}(1-SP_{x0})$. In contrast, if independence is assumed while with preserving differentiality on both variables, $\pi_{11}^* = SE_{y1}\pi_{11}SE_{x1} + SE_{y0}\pi_{01}(1-SP_{x1}) + (1-SP_{y1})\pi_{10}SE_{x0} + (1-SP_{y0})\pi_{00}(1-SP_{x0})$. Under the most simplified situation, assuming independence and nondifferentiality at the same time, $\pi_{11}^* = SE_y\pi_{11}SE_x + SE_{y1}\pi_{01}(1-SP_x) + (1-SP_y)\pi_{10}SE_x + (1-SP_y)\pi_{00}(1-SP_x)$. Other π^*s can be derived similarly.

4.1.3 Generalized Matrix Method

We generalize the concept of the matrix method and its extensions (1, 28) by allowing flexible incorporation of various situations. In general, one is able to relate surrogate and true cell probabilities via the equality $\mathbf{\Pi}^* = \boldsymbol{A}\mathbf{\Pi}$, where $\mathbf{\Pi} = (\pi_{11} \quad \pi_{01} \quad \pi_{10} \quad \pi_{00})'$, $\mathbf{\Pi}^* = (\pi_{11}^* \quad \pi_{01}^* \quad \pi_{10}^* \quad \pi_{00}^*)'$ and the definition of $\mathbf{A}$ varies according to the assumptions made. For differential and dependent misclassification, $\mathbf{A}$ takes its most general from:

$$\boldsymbol{A} = \begin{bmatrix} SE_{y11}SE_{x1} & SE_{y10}(1-SP_{x1}) & (1-SP_{y11})SE_{x0} & (1-SP_{y10})(1-SP_{x0}) \\ SE_{y01}(1-SE_{x1}) & SE_{y00}SP_{x1} & (1-SP_{y01})(1-SE_{x0}) & (1-SP_{y00})SP_{x0} \\ (1-SE_{y11})SE_{x1} & (1-SE_{y10})(1-SP_{x1}) & SP_{y11}SE_{x0} & SP_{y10}(1-SP_{x0}) \\ (1-SE_{y01})(1-SE_{x1}) & (1-SE_{y00})SP_{x1} & SP_{y01}(1-SE_{x0}) & SP_{y00}SP_{x0} \end{bmatrix}$$

If instead assuming traditional differential misclassification with independence,

$$\boldsymbol{A} = \begin{bmatrix} SE_{y1}SE_{x1} & SE_{y0}(1-SP_{x1}) & (1-SP_{y1})SE_{x0} & (1-SP_{y0})(1-SP_{x0}) \\ SE_{y1}(1-SE_{x1}) & SE_{y0}SP_{x1} & (1-SP_{y1})(1-SE_{x0}) & (1-SP_{y0})SP_{x0} \\ (1-SE_{y1})SE_{x1} & (1-SE_{y0})(1-SP_{x1}) & SP_{y1}SE_{x0} & SP_{y0}(1-SP_{x0}) \\ (1-SE_{y1})(1-SE_{x1}) & (1-SE_{y0})SP_{x1} & SP_{y1}(1-SE_{x0}) & SP_{y0}SP_{x0} \end{bmatrix}$$

which has the same form as defined by Greenland et al. (28).

Under the circumstance of nondifferential and independent misclassification,

$$\boldsymbol{A} = \begin{bmatrix} SE_{y}SE_{x} & SE_{y}(1-SP_{x}) & (1-SP_{y})SE_{x} & (1-SP_{y})(1-SP_{x}) \\ SE_{y}(1-SE_{x}) & SE_{y}SP_{x} & (1-SP_{y})(1-SE_{x}) & (1-SP_{y})SP_{x} \\ (1-SE_{y})SE_{x} & (1-SE_{y})(1-SP_{x}) & SP_{y}SE_{x} & SP_{y}(1-SP_{x}) \\ (1-SE_{y})(1-SE_{x}) & (1-SE_{y})SP_{x} & SP_{y}(1-SE_{x}) & SP_{y}SP_{x} \end{bmatrix}$$

and with some algebraic work, one can easily show that this equation is equivalent to that underlying Barron's original matrix method (1). With algebraic work, it can be shown that A is invertible if and only if $SE_x + SP_x - 1 \neq 0$ and $SE_y + SP_y - 1 \neq 0$. Under usual circumstances, the chance of correctly classifying a diagnosis should be greater than a random chance; thus, in realistic setting it should be always the case that $SE_x + SP_x - 1 > 0$ and $SE_y + SP_y - 1 > 0$.

Thus, the generalized matrix method can be derived immediately as $\mathbf{\Pi}=\mathbf{A}^{-1}\mathbf{\Pi}^{*}$.

4.1.4 Generalized Inverse Matrix Method

The inverse matrix method directly expresses true cell probabilities as sums of products of surrogate cell probabilities and predictive values, without inversion computations involved. Here, we expand Marshall's inverse matrix method (8) to a general context when both variables are misclassified in a 2×2 table. For example, by laws of probability,

$\pi_{11} = PPV_{y11}\pi^{*}_{11}PPV_{x1} + (1-NPV_{y11})\pi^{*}_{10}PPV_{x0} + PPV_{y01}\pi^{*}_{01}(1-NPV_{x1}) + (1-NPV_{y01})\pi^{*}_{00}(1-NPV_{x0})$ under dependent and differential misclassification.

Packaging linear equations into matrices, the form of generalized inverse matrix method is the same as Marshall's original proposal of **Π**=**BΠ***. However, in our approach, the matrix **B** takes a more complicated form in characterizing misclassification in both the X and Y variables:

$$\boldsymbol{B} = \begin{bmatrix} PPV_{y11}PPV_{x1} & PPV_{y10}(1-NPV_{x1}) & (1-NPV_{y11})SE_{x0} & (1-NPV_{y10})(1-NPV_{x0}) \\ PPV_{y01}(1-PPV_{x1}) & PPV_{y00}NPV_{x1} & (1-NPV_{y01})(1-PPV_{x0}) & (1-NPV_{y00})NPV_{x0} \\ (1-PPV_{y11})PPV_{x1} & (1-PPV_{y10})(1-NPV_{x1}) & NPV_{y11}PPV_{x0} & NPV_{y10}(1-NPV_{x0}) \\ (1-PPV_{y01})(1-PPV_{x1}) & (1-PPV_{y00})NPV_{x1} & NPV_{y01}(1-PPV_{x0}) & NPV_{y00}NPV_{x0} \end{bmatrix}$$

In contrast to the generalized matrix method, there is no matrix inversion involved in computing the corrected OR through the generalized inverse matrix method.

4.1.5 Estimation of Misclassification Probabilities and Variance

Our primary measure of association of interest is the OR, and the estimate of the corrected OR is $\widehat{OR} = \widehat{\pi_{11}}\widehat{\pi_{00}} \Big/ \widehat{\pi_{01}}\widehat{\pi_{00}}$. In all of the approaches presented above, estimation of misclassification probabilities is crucial. Although misclassification probability estimates from external studies may be used, there is always good reason to suspect that they may vary from study to study. Thus, when possible we recommend the use of an internal validation subsample randomly selected from one's current study, in which true disease and exposure status is measured using gold standard methods. The primary appeal of adopting internal validation sampling is the potential of avoiding assuming "transportability" in misclassification probabilities.

Table 4.1 Description and likelihood contributions for 16 possible types of observations under the internal validation sampling[a].

Obs. Type	Description	Likelihood contribution in terms of SE and SP	Likelihood contribution in terms of Predictive Values
1	$X^*=1,\ Y^*=1,\ X=1,Y=1$	$SE_{y11}SE_{x1}\pi_{11}$	$PPV_{y11}PPV_{x1}\ \pi_{11}^*$
2	$X^*=1,\ Y^*=1,\ X=1,Y=0$	$(1-SP_{y11})SE_{x0}\pi_{10}$	$(1-PPV_{y11})PPV_{x1}\ \pi_{11}^*$
3	$X^*=1,\ Y^*=1,\ X=0,Y=1$	$SE_{y01}(1-SP_{x1})\pi_{01}$	$PPV_{y01}(1-PPV_{x1})\ \pi_{11}^*$
4	$X^*=1,\ Y^*=1,\ X=0,Y=0$	$(1-SP_{y01})(1-SP_{x0})\pi_{00}$	$(1-PPV_{y01})(1-PPV_{x1})\ \pi_{11}^*$
5	$X^*=1,\ Y^*=0,\ X=1,Y=1$	$(1-SE_{y11})SE_{x1}\pi_{11}$	$(1-NPV_{y11})PPV_{x0}\ \pi_{10}^*$
6	$X^*=1,\ Y^*=0,\ X=1,Y=0$	$SP_{y11}SE_{x0}\pi_{10}$	$NPV_{y11}PPV_{x0}\ \pi_{10}^*$
7	$X^*=1,\ Y^*=0,\ X=0,Y=1$	$(1-SE_{y01})(1-SP_{x1})\pi_{01}$	$(1-NPV_{y01})(1-PPV_{x0})\ \pi_{10}^*$
8	$X^*=1,\ Y^*=0,\ X=0,Y=0$	$SP_{y01}\ (1-SP_{x0})\pi_{00}$	$NPV_{y01}(1-PPV_{x0})\ \pi_{10}^*$
9	$X^*=0,\ Y^*=1,\ X=1,Y=1$	$SE_{y11}(1-SE_{x1})\pi_{11}$	$PPV_{y10}(1-NPV_{x1})\ \pi_{01}^*$
10	$X^*=0,\ Y^*=1,\ X=1,Y=0$	$(1-SP_{y11})(1-SE_{x0})\pi_{10}$	$(1-PPV_{y10})\ (1-NPV_{x1})\ \pi_{01}^*$
11	$X^*=0,\ Y^*=1,\ X=0,Y=1$	$SE_{y01}SP_{x1}\pi_{01}$	$PPV_{y00}NPV_{x1}\pi_{01}^*$
12	$X^*=0,\ Y^*=1,\ X=0,Y=0$	$(1-SP_{y01})SP_{x0}\pi_{00}$	$(1-\ PPV_{y00})\ NPV_{x1}\pi_{01}^*$
13	$X^*=0,\ Y^*=0,\ X=1,Y=1$	$(1-SE_{y11})(1-SE_{x1})\pi_{11}$	$(1-NPV_{y10})\ (1-NPV_{x0})\ \ \pi_{00}^*$
14	$X^*=0,\ Y^*=0,\ X=1,Y=0$	$SP_{y11}\ (1-SE_{x0})\pi_{10}$	$NPV_{y10}(1-NPV_{x0})\pi_{00}^*$
15	$X^*=0,\ Y^*=0,\ X=0,Y=1$	$(1-SE_{y01})SP_{x1}\pi_{01}$	$(1-NPV_{y00})NPV_{x0}\pi_{00}^*$
16	$X^*=0,\ Y^*=0,\ X=0,Y=0$	$SP_{y01}SP_{x0}\pi_{00}$	$NPV_{y00}NPV_{x0}\pi_{00}^*$

a. See Section 4.1.1 foe definitions of terms.

When allowing full generality, i.e., dependent and differential misclassification, it can be shown that the likelihood approach is equivalent when parameterized in terms of predictive values and SE/SP. There are in total 16 types of validation observations if validations on X and Y are measured simultaneously for each subject of the subsample, and Table 4.1 shows the representations from two approaches. The main likelihood is always written as

$$L_m = \prod_{i=1}^{n_m} \pi_{11}^{*(y_i^* x_i^*)} \pi_{01}^{*((1-x_i^*)y_i^*)} \pi_{10}^{*(x_i^*(1-y_i^*))} \pi_{00}^{*((1-x_i^*)(1-y_i^*))},$$

while if expressed in terms of SE and SP, all the π^*s can be further written out (see Section 4.1.2).

The internal validation subsample likelihood is $L_v = \prod_{j=1}^{16} L_{v_j}^{n_{vj}}$, where L_{vj} is the likelihood term read from type j in Table 1, while n_{vj} is the total number of observations in the jth type (j=1,2,…16). Note that the total validation sample size $n_v = \sum_{j=1}^{16} n_{vj}$.The overall likelihood is proportional to $L_m \times L_v$. There are no closed-form solutions for the MLEs based on the complete likelihood written in terms of SE and SP, but closed-forms exist for the version expressed in terms of predictive values. For example, one can derive $\widehat{PPV_{y11}} = \frac{\sum_{i=1}^{n_v} I_{y_i=1, x_i=1, x^*{}_i=1}}{n_v}$. Since under the circumstance of dependent and differential misclassification, the two paramerizations are equivalent, we may obtain MLEs for the $\widehat{SE}$ and $\widehat{SP}$ parameters as functions of $\widehat{PPV}$'s and $\widehat{NPV}$'s. For example,

$$\widehat{SE_{y11}} = \frac{\widehat{\mathrm{PPV}_{\mathrm{y11}}}\,\widehat{\mathrm{PPV}_{\mathrm{x1}}}\,\widehat{\pi_{11}^*} + \widehat{\mathrm{PPV}_{\mathrm{y10}}}(1 - \widehat{\mathrm{NPV}_{\mathrm{x1}}})\widehat{\pi_{01}^*}}{\widehat{\pi_{11}}}$$

However, when misclassification is non-differential or differential but independent, the equivalence does not hold any more, since nondifferentiality under one parameterization poses nonlinear constraints on the parameters of the other. In such cases there is no simple closed-form for the $\widehat{SE}$'s, $\widehat{SP}$'s and $\hat{\pi}$'s. Therefore, if supplying the generalized matrix method with crude estimates of SE and SP parameters based on simple corresponding sample proportions, the corrected $\widehat{OR}$ will not be as efficient as the MLE. These conclusions are consistent with previous findings, though under a simpler and slightly different context (42).

In general, we recommend the use of the ML approach in the interest of optimal efficiency and based on the ease of computing standard errors. Optimizing the likelihood in both parameterization paths is readily available by taking advantage of numerical procedures in standard statistical software. We view matrix and inverse matrix forms as a convenient identity-based methods to compute the corrected $\widehat{OR}$. Through tedious but straightforward multivariate delta-method practices, the approximate standard error of the corrected $\ln(\widehat{OR})$ based on the general matrix and inverse matrix methods could also be computed.

4.1.6 Notes on Case-Control Studies

Though throughout the focus has been on cross-sectional sampling, case-control, as an important sampling scheme in epidemiology, is also worth discussion. We consider "case-control" studies as those where case over-sampling is conducted based on the error-prone responses. In other words, observations with $Y^*=1$ (cases) are sampled with a greater probability than those with $Y^*=0$ (controls). Thus, (mis)classification occurs before the case-control sampling. Prior work (26) has noticed that supplying the

population misclassification probabilities to the correction methods will not yield valid estimates; however, with nondifferential misclassification, the validity of the analytic results could be restored by introducing the sampling fraction between cases and controls into the correction. With straightforward algebraic work, we can make the following definitions:

$$\text{“Operating” SE} = \frac{\rho_1 \times SE}{\rho_1 \times SE + \rho_0 \times (1 - SE)}$$

$$\text{“Operating” SP} = \frac{\rho_0 \times SP}{\rho_0 \times SP + \rho_1 \times (1 - SP)}$$

where ρ_i=Pr(being selected | Y^*=i).

Lyles *et al.* (16) further examined the impact of case oversampling on correcting outcome misclassification in ordinary logistic regressions. They pointed out that the “operating” misclassification probabilities under the “case-control” sampling differ from the population diagnostic properties, and that this difference is essential to validly estimating the corrected ORs. For this reason, the main/internal validation study is favorable because it readily permits estimating “operating” misclassification probabilities. With empirical evidence and some analytic work, they also suggested that the corrected OR in case-control studies is generally valid with proper handling of the analysis, when misclassification is nondifferential. However, the validity only exists for special cases under differential misclassification.

With oversampling of “cases” (Y^*=1), the method described in the previous sections yields valid estimation of the OR, assuming misclassification is nondifferential. As observed in (16), with such oversampling, the estimate of SE_y is inflated while SP_y is

deflated, reflecting the "operating" characteristics of the diagnostic method. In contrast, SE_x and SP_x are not affected. The difference between "operating" and population properties reinforces the importance of incorporating an internal validation data. Otherwise, one would have to know the corresponding selection probabilities in order to reasonably convert population SE/SP to "operating" SE/SP. However, when the nondifferential misclassification assumption is not met, the validity of the estimated OR based on the main/internal validation design does not hold. A brief argument can be made as follows. Consider an ordinary logistic model to specify the relationship between Y and X as

$logit\{\Pr(Y=1|X)\} = \beta_0 + \beta_1 x$ where $\exp(\beta_1)$=OR.

Taking the dependent and differential misclassification as an example,

$$\log\left\{\frac{\Pr(Y=1|X=1,S=1)}{\Pr(Y=0|X=1,S=1)}\right\}$$

$$= \beta_0 + \beta_1$$

$$+ \log\left(\frac{\rho_0(1-SE_{y10})(1-SE_{x1}) + \rho_0(1-SE_{y11})SE_{x1} + \rho_1 SE_{y10}(1-SE_{x1}) + \rho_1 SE_{y11}SE_{x1}}{\rho_0 SP_{y10}(1-SE_{x0}) + \rho_0 SP_{y11}SE_{x0} + \rho_1(1-SP_{y10})(1-SE_{x0}) + \rho_1(1-SP_{y11})SE_{x0}}\right) \quad (4.6)$$

where S=(0,1) with 1 if selected. Similarly,

$$\log\left\{\frac{\Pr(Y=1|X=0,S=1)}{\Pr(Y=0|X=0,S=1)}\right\}$$

$$= \beta_0 + \log\left(\frac{\rho_0(1-SE_{y00})SP_{x1} + \rho_0(1-SE_{y01})(1-SP_{x1}) + \rho_1 SE_{y00}SP_{x1} + \rho_1 SE_{y01}(1-SP_{x1})}{\rho_0 SP_{y00}SP_{x0} + \rho_0 SP_{01}(1-SP_{x0}) + \rho_1(1-SP_{y00})SP_{x0} + \rho_1(1-SP_{y01})(1-SP_{x0})}\right) \quad (4.7)$$

where ρ_i=Pr(being selected | Y^*=i). It is assumed here that ρ_i only depends on the status of Y^*. In other words, Pr(being selected | Y^*=i, Y=y, X=x, X^*=x^*)= Pr(being selected | Y^*=i), indicating completely random case-dependent sampling.

Thus, it is clear that the logistic model does not hold any more, and the OR is a function of β, and the SE's and SP's associated with X and Y. When nondifferential misclassification is assumed,

$$logit\{\Pr(Y = 1|X, S = 1)\} = \beta_0 + \frac{\rho_1}{\rho_0} + \beta_1 x$$

where the logistic model reserves and the selection probabilities are absorbed into the intercept. Interestingly, as long as Y is nondifferentially misclassified, even while X is differentially misclassified, one can easily show with eqn.s (4.6) and (4.7) that the logistic regression still holds, and the estimate of the OR is valid. Similarly, if the case-control sampling is conducted on the basis of X^*, logistic regression holds regardless of the misclassification mechanism of Y. Note here again that the selection probability is assumed only dependent on the status of Y^*. For example, $\rho_i = \Pr(S = 1|Y^* = i) = \Pr(S = 1|Y^* = i, X = x, X^* = x^*)$ i = 0,1.

4.1.7 Model Selection

When correcting the estimate of the OR, we would ideally choose the misclassification mechanism fitting best with the data. As described in Section 4.1.1.4, there are other types of misclassification models than those mentioned in Section 4.1.1.1 through Section 4.1.1.3. Here we provide a straightforward model selection procedure to guide practitioners to pick the desired model. For the ease of discussion, denote the dependent and differential misclassification model as "model 1", followed by "model 2" (the independent and differential misclassification model in Section 4.1.1.2), "model 3" (the model with differential X^* and nondifferential Y^*), "model 4" (the model with nondifferential X^* and differential Y^*) and "model 5" (the completely nondifferential

model in Section 4.1.1.3). Denote $\widehat{OR}_i=\widehat{OR}$ from Model i and LR_i=-2log Likelihood of Model i (i=1,…,5). In practice, one may use a 2-step procedure to implement model selection.

Step 1: By treating $\widehat{OR}_1$ as a standard, compute the relative change in $\widehat{OR}$ as $\left|\frac{\widehat{OR}_i-\widehat{OR}_1}{\widehat{OR}_1}\right|$. (i=2,3,4,5). Pick models with the relative change less than θ, a pre-specified threshold. For example, one may specify θ=0.1.

Step 2: Among those models picked in Step 1, perform the likelihood ratio test (LRT) on model i with the smallest relative change to test H_0: Model reduction is appropriate. LR test statistic=$-2(LR_i-LR_1)$, which under H_0 follows a χ^2 distribution with a corresponding df. For example, if i=2, df=4. When i=3 or 4, df=6, and df=8 if i=5.

The performance of the proposed model selection strategy is evaluated via simulations in Section 4.2.3. It has also been applied to the real data example.

4.1.8 Comments Regarding Null Testing

With complete nondifferentiality in misclassification of both X and Y (Model 5), one can show that OR=1 is a necessary and sufficient condition for OR^*=1, suggesting that the hypothesis test of no association is still valid though with lowered power. Thus, with nondifferentiality, if one's interest is only in hypothesis testing, the naïve analysis is defensible.

We show below that $OR = 1 \Leftrightarrow OR^* = 1$ when H_0 is true.

1) $\boldsymbol{OR = 1 \Rightarrow OR^* = 1}$

Define $P_i = P(Y = 1|X = i)$; $P_x = P(X = 1)$; $P_i^* = P(Y^* = 1|X^* = i)$; $P_x^* = P(X^* = 1)$, and note that

$$OR = 1 \Leftrightarrow P_1 = P_0 = P.$$

Assuming nondifferentiality, we have

$$\pi_{11}^* = SE_y\pi_{11}SE_x + (1 - SP_y)\pi_{10}SE_x + SE_y\pi_{01}(1 - SP_x) + (1 - SP_y)\pi_{00}(1 - SP_x)$$

Rewriting in terms of P and P_x, this becomes

$$\pi_{11}^* = SE_y SE_x PP_x + (1 - SP_y)SE_x(1 - P)P_x + SE_y(1 - SP_x)P(1 - P_x) + (1 - SP_y)(1 - SP_x)(1 - P)(1 - P_x)$$

$$= [SE_yP + (1 - SP_y)(1 - P)][SE_xP_x + (1 - SP_x)(1 - P_x)].$$

Similarly, we can derive that

$$\pi_{10}^* = SE_y(1 - SE_x)PP_x + (1 - SP_y)(1 - SE_x)(1 - P)P_x + SE_y SP_x P(1 - P_x) + (1 - SP_y)SP_x(1 - P)(1 - P_x)$$

$$= [(1 - SE_y)P + SP_y(1 - P)][SE_xP_x + (1 - SP_x)(1 - P_x)],$$

$$\pi_{01}^* = (1 - SE_y)SE_x PP_x + SP_y SE_x(1 - P)P_x + (1 - SE_y)(1 - SP_x)P(1 - P_x) + SP_y(1 - SP_x)(1 - P)(1 - P_x)$$

$$= [SE_yP + (1 - SP_y)(1 - P)][(1 - SE_x)P_x + SP_x(1 - P_x)],$$

and

$$\pi_{00}^* = (1 - SE_y)(1 - SE_x)PP_x + SP_y(1 - SE_x)(1 - P)P_x + (1 - SE_y)SP_x P(1 - P_x) + SP_y SP_x(1 - P)(1 - P_x)$$

$$= [(1 - SE_y)P + SP_y(1 - P)][(1 - SE_x)P_x + SP_x(1 - P_x)].$$

Note that to show $OR^* = 1$ is equivalent to showing $P_1^*/P_0^* = 1$.

Since $P_1^* = \frac{\pi_{11}^*}{\pi_{11}^*+\pi_{10}^*} = \frac{[SE_y P+(1-SP_y)(1-P)][SE_x P_x+(1-SP_x)(1-P_x)]}{[SE_x P_x+(1-SP_x)(1-P_x)]} = [SE_y P + (1-SP_y)(1-P)]$

and $P_0^* = \frac{\pi_{01}^*}{\pi_{01}^*+\pi_{00}^*} = \frac{[SE_y P+(1-SP_y)(1-P)][(1-SE_x)P_x+SP_x(1-P_x)]}{[(1-SE_x)P_x+SP_x(1-P_x)]} = [SE_y P + (1-SP_y)(1-P)]$

$P_1^*/P_0^* = 1 \Leftrightarrow OR^* = 1$

Therefore, $\boldsymbol{OR = 1 \Rightarrow\ OR^* = 1}$.

2) $\boldsymbol{OR = 1 \Leftarrow\ OR^* = 1}$

Note that $OR^* = 1 \Leftrightarrow P_1^*/P_0^* = 1$ where $P_i{}^* = P(Y^* = 1|X^* = i) = P^*$ and $P_x{}^* =$

$P(X^* = 1)$

Solving for π_{ij} from the linear system above, we get

$$\pi_{11} = \frac{(SP_x-1)(SP_y-1)\pi_{00}^* + (SP_x-1)SP_y\pi_{01}^* + SP_x(SP_y-1)\pi_{10}^* + SP_xSP_y\pi_{11}^*}{(SE_x+SP_x-1)(SE_y+SP_y-1)},$$

$$\pi_{10} = \frac{(SP_x-1)SE_y\pi_{00}^* + (SP_x-1)(SE_y-1)\pi_{01}^* + SP_xSE_y\pi_{10}^* + SP_x(SE_y-1)\pi_{11}^*}{(SE_x+SP_x-1)(SE_y+SP_y-1)},$$

$$\pi_{01} = \frac{SE_x(SP_y-1)\pi_{00}^* + SE_xSP_y\pi_{01}^* + (SE_x-1)(SP_y-1)\pi_{10}^* + (SE_x-1)SP_y\pi_{11}^*}{(SE_x+SP_x-1)(SE_y+SP_y-1)}, \text{ and}$$

$$\pi_{00} = \frac{SE_xSE_y\pi_{00}^* + SE_x(SE_y-1)\pi_{01}^* + (SE_x-1)SE_y\pi_{10}^* + (SE_x-1)(SE_y-1)\pi_{11}^*}{(SE_x+SP_x-1)(SE_y+SP_y-1)}.$$

Replacing the π^*s above with the appropriate functions of P^* and $P_x{}^*$, we can rewrite the expressions as follows:

$$\pi_{11} = \frac{(SP_x-1)(SP_y-1)(1-P^*)(1-P_x^*)+(SP_x-1)SP_yP^*(1-P_x^*)+SP_x(SP_y-1)(1-P^*)P_x^*+SP_xSP_yP^*P_x^*}{(SE_x+SP_x-1)(SE_y+SP_y-1)},$$

$$\pi_{10} = \frac{(SP_x-1)SE_y(1-P^*)(1-P_x^*)+(SP_x-1)(SE_y-1)P^*(1-P_x^*)+SP_xSE_y(1-P^*)P_x^*+SP_x(SE_y-1)P^*P_x^*}{(SE_x+SP_x-1)(SE_y+SP_y-1)},$$

$$\pi_{01} = \frac{SE_x(SP_y-1)(1-P^*)(1-P_x^*)+SE_xSP_yP^*(1-P_x^*)+(SE_x-1)(SP_y-1)(1-P^*)P_x^*+(SE_x-1)SP_yP^*P_x^*}{(SE_x+SP_x-1)(SE_y+SP_y-1)}, \text{ and}$$

$$\pi_{00} = \frac{SE_xSE_y(1-P^*)(1-P_x^*)+SE_x(SE_y-1)P^*(1-P_x^*)+(SE_x-1)SE_y(1-P^*)P_x^*+(SE_x-1)(SE_y-1)P^*P_x^*}{(SE_x+SP_x-1)(SE_y+SP_y-1)}.$$

It follows that

$$P_1 = \frac{\pi_{11}}{\pi_{11}+\pi_{10}} = \frac{(SP_x+P_x^*-1)(SP_y+P^*-1)}{(SE_y+SP_y-1)(SP_x+P_x^*-1)} = \frac{SP_y+P^*-1}{SE_y+SP_y-1}, \text{ and}$$

$$P_0 = \frac{\pi_{01}}{\pi_{01}+\pi_{00}} = \frac{(SE_x-P_x^*)(SP_y+P^*-1)}{(SE_y+SP_y-1)(SE_x-P_x^*)} = \frac{SP_y+P^*-1}{SE_y+SP_y-1}.$$

Thus, $P_1/P_0 = 1 \Longrightarrow OR = 1.$

Therefore, we have established that $\boldsymbol{OR = 1 \Leftarrow OR^* = 1.}$

Combining 1) and 2), we conclude that the equivalence relationship holds, i.e., $\boldsymbol{OR = 1 \Rightarrow OR^* = 1}$. This result is reminiscent of classic findings in the case of single variable misclassification (19).

4.2. SIMULATION STUDIES

4.2.1 Study I: Mimicking Real-data Example

Our first simulation experiment evaluates the performance of our methods under conditions mimicking the HERS example (Section 4.3). The cell counts are simulated from a multinomial distribution with cell probabilities and main and internal validation sample sizes similar to those observed from the HERS example. Error-prone response Y^* and exposure X^* are generated with misclassification probabilities estimated from the HERS sample. The underlying misclassification process is assumed dependent and differential. For each of 500 simulated datasets, we conduct naïve analysis associating Y^* with X^*, true analysis with Y and X, and main/internal validation analyses based on Models 1 through 5.

Table 4.2 summarizes the results. The naïve analysis yields a biased result away from the null. Model 1 produces the corrected OR estimate closest to the gold standard OR, with tolerable sacrifice in efficiency. The 95% CI coverage of Model 1 is also excellent. When reducing Model 1 to other simpler versions, by assuming independence or nondifferentiality, the results are biased, indicating the reduced models are not consistent with the data generation process. Note that with the simplest model assuming nondifferential misclassification (Model 5), the corrected result is similar to the naïve result (in fact, arguably worse), suggesting the importance of a careful model selection.

The corrected results using the generalized matrix methods discussed in Section 4.1.1.1 agree well with the MLEs, when ML estimates of misclassification probabilities are supplied. However, when the simpler crude estimates obtained from the validation

subsample are inserted into the generalized matrix method, the results are not satisfying, even producing negative estimates of probabilities in some cases (results not shown). Thus, in practice, we favor the ML approach in order to obtain both valid and efficient corrections.

Table 4.2 Results of Simulations Addressing Main/Internal Validation Study-Based Analysis Mimicking HERS Data[a.b]

Model	**$\log(\widehat{OR})$ (SD)**	**95% CI Coverage**
Naive[c]	1.42(0.23)	67.4%
Gold Standard[d]	1.15(0.18)	93.6%
Model 1[e]	1.16(0.34)	95.7%
Model 2[f]	1.28(0.34)	93.3%
Model 3[g]	1.41(0.33)	57.8%
Model 4[h]	1.35(0.33)	89.0%
Model 5[i]	1.58(0.31)	72.4%

a. 500 simulation studies; 229 internal validation observations and 687 main study observations per simulation. b. True ln(OR)=1.14. c. $\widehat{OR}$ is calculated using Y^* and X^* data. d. $\widehat{OR}$ is calculated from using Y and X data. e. Model is fitted assuming dependent and differential misclassification. f. Model is fitted assuming independent and differential misclassification. g. Model is fitted assuming differential misclassification in Y and nondifferential misclassification in X. h. Model is fitted assuming nondifferential misclassification in Y and differential misclassification in X. i. Model is fitted assuming completely nondifferential misclassification.

4.2.2 Study II: Different Types of Misclassification

Tables 4.3-4.6 summarize the results when the true underlying misclassification mechanism is either independent and differential (Model 2), differential for Y and nondifferential for X (Model 3), nondifferential for Y and differential for X (Model 4), or completely nondifferential (Model 5). In this experiment, 500 simulated datasets are generated accordingly. Experiments under the various misclassification mechanisms demonstrate that the proposed approach performs quite well under different situations. For example, Table 4.5 summarizes the results when only X is differentially misclassified. Without correction, the naïve analytic approach produces ln(OR) estimates that are biased and on the wrong side of the null on average. This contradiction reinforces the importance of appropriate correction. Model 1 is the fully general misclassification model; thus, the corrected estimate agrees well with the true value. Similarly, though Model 2 is unnecessarily complicated and further model reduction should be appropriate, Model 2 is still valid. Models 3 and Model 5 are not correct models for this example; therefore, results based on them are noticeably biased. Especially, the corrected result of Model 5 is biased toward the null by ~30%, compared to the true value, reinforcing the notion that when the misclassification assumed is too simplistic, validity is lost. In the meantime, the efficiency loss fitting overly general models appear small.

Table 4.3 Results of Simulations Addressing Main/Internal Validation Study-Based Analysis with Model 2 as the True Underlying Model[a.b]

Model	**$\log(\widehat{OR})$ (SD)**	**95% CI Coverage**
Naive[c]	0.66(0.25)	67.4%
Gold Standard[d]	1.12(0.18)	93.6%
Model 1[e]	1.12(0.34)	96.0%
Model 2[f]	1.09(0.34)	94.6%
Model 3[g]	1.03(0.33)	90.3%
Model 4[h]	1.17(0.33)	91.2%
Model 5[i]	1.03(0.33)	89.7%

a. 500 simulation studies; 229 internal validation observations and 687 main study observations per simulation. b. True ln(OR) =1.10. c. $\widehat{OR}$ is calculated using Y^* and X^* data. d. $\widehat{OR}$ is calculated from using Y and X data. e. Model is fitted assuming dependent and differential misclassification. f. Model is fitted assuming independent and differential misclassification. g. Model is fitted assuming differential misclassification in Y and nondifferential misclassification in X. h. Model is fitted assuming nondifferential misclassification in Y and differential misclassification in X. i. Model is fitted assuming completely nondifferential misclassification.

Table 4.4 Results of Simulations Addressing Main/Internal Validation Study-Based Analysis with Model 3 as the True Underlying Model[a.b]

Model	$\log(\widehat{OR})$ (SD)	95% CI Coverage
Naive[c]	0.61(0.15)	22.8%
Gold Standard[d]	1.00(0.14)	94.8%
Model 1[e]	1.01(0.26)	97.6%
Model 2[f]	0.99(0.26)	95.0%
Model 3[g]	1.01(0.25)	95.6%
Model 4[h]	1.13(0.25)	90.4%
Model 5[i]	1.20(0.24)	85.0%

a. 500 simulation studies; 229 internal validation observations and 687 main study observations per simulation. b. True ln(OR)=1. c. $\widehat{OR}$ is calculated from using Y^* and X^* data. d. $\widehat{OR}$ is calculated from using Y and X data. e. Model is fitted assuming dependent and differential misclassification. f. Model is fitted assuming independent and differential misclassification. g. Model is fitted assuming differential misclassification in Y and nondifferential misclassification in X. h. Model is fitted assuming nondifferential misclassification in Y and differential misclassification in X. i. Model is fitted assuming completely nondifferential misclassification.

Table 4.5 Results of Simulations Addressing Main/Internal Validation Study-Based Analysis with Model 4 as the True Underlying Model[a.b]

Model	$\log(\widehat{OR})$ (SD)	95% CI Coverage
Naive[c]	-0.13(0.15)	0
Gold Standard[d]	1.01(0.14)	95.0%
Model 1[e]	1.00(0.26)	95.0%
Model 2[f]	0.99(0.25)	94.8%
Model 3[g]	0.78(0.25)	82.6%
Model 4[h]	0.99(0.25)	94.8%
Model 5[i]	0.74(0.25)	76.4%

a. 500 simulation studies; 229 internal validation observations and 687 main study observations per simulation. b. True ln (OR) =1. c. $\widehat{OR}$is calculated from using Y^* and X^* data. d. $\widehat{OR}$is calculated from using Y and X data. e. Model is fitted assuming dependent and differential misclassification. f. Model is fitted assuming independent and differential misclassification. g. Model is fitted assuming differential misclassification in Y and nondifferential misclassification in X. h. Model is fitted assuming nondifferential misclassification in Y and differential misclassification in X. i. Model is fitted assuming completely nondifferential misclassification.

Table 4.6 Results of Simulations Addressing Main/Internal Validation Study-Based Analysis with Model 5 as the True Underlying Model[a.b]

Model	$\log(\widehat{OR})$ (SD)	95% CI Coverage
Naive[c]	0.46(0.27)	40.0%
Gold Standard[d]	1.06(0.18)	93.8%
Model 1[e]	1.05(0.33)	97.6%
Model 2[f]	1.06(0.32)	96.0%
Model 3[g]	1.06(0.32)	95.8%
Model 4[h]	1.06(0.32)	96.0%
Model 5[i]	1.06(0.31)	96.2%

a. 500 simulation studies; 229 internal validation observations and 687 main study observations per simulation. b. True ln(OR) =1.05. c. $\widehat{OR}$is calculated from using Y^* and X^* data. d. $\widehat{OR}$is calculated from using Y and X data. e. Model is fitted assuming dependent and differential misclassification. f. Model is fitted assuming independent and differential misclassification. g. Model is fitted assuming differential misclassification in Y and nondifferential misclassification in X. h. Model is fitted assuming nondifferential misclassification in Y and differential misclassification in X. i. Model is fitted assuming completely nondifferential misclassification.

4.2.3 Study III: Performance of Model Selection

Results in Sections 4.2.1 and 4.2.2 suggest the importance of a careful model selection to ensure the model is specified correctly (or, at least, generally enough) and the results are valid. By adopting the model selection procedure described in Section 4.1.7, 92.8% of the time we correctly select Model 1 out of 500 simulations mimicking the HERS example, yielding a valid corrected result. The 95% CI coverage is also satisfactory (Table 4.7). When the true underlying model is the one with independent and differential/nondifferential mechanisms, similar conclusions are observed. For example, when the true model has both X and Y are nondifferentially misclassified (Table 4.8), by selecting the model, the validity of the analysis is maintained, A small efficiency gain is also observed, based on modele selection, as opposed to fitting overly general models.

In practice, since Model 1 is always valid, we suggest that the users consider adopting Model 1 all the time when there is a rich validation resource. Only if users strongly aim to obtain a more precise confidence interval, given that the validity is ensured, we recommend a careful model selection in order to improve the efficiency.

Table 4.7 Performance of Model Selection with Main/Internal Validation Study-Based Analysis Mimicking HERS Data[a,b]

Model	**$\log(\widehat{OR})$ (SD)**	**Mean SE**	**95% CI Coverage**
Naïve	1.42(0.23)	0.23	67.4%
Gold	1.15(0.18)	0.18	93.6%
Model 1	1.17(0.35)	0.34	95.6%
Model Selection Result (α=0.05[c],θ=0.1)	1.18(0.36)	0.24	94.8%

Percentage of runs for which model was selected				
Pr(General)	Pr(diff X and Y)	Pr (diff Y nondiff X)	Pr (nondiff Y diff X)	Pr(nondiff)
78.0%	9.0%	2.0%	11.2%	0%

a. 500 simulation studies; 229 internal validation observations and 687 main study observations per simulation. b. True ln(OR)=1.14. c. α is the significance level used for LRT.

Table 4.8 Performance of Model Selection with Main/Internal Validation Study-Based Analysis under Completely Nondifferential Model[a,b]

Model	**$\log(\widehat{OR})$ (SD)**	**Mean SE**	**95% CI Coverage**
Naïve	0.46(0.27)	0.26	40.0%
Gold	1.06(0.18)	0.17	93.8%
Model 1	1.05 (0.34)	0.34	97.5%
Model 5	1.06(0.31)	0.31	96.2%
Model Selection Result (α=0.05[c], θ=0.1)	1.06(0.31)	0.32	96.0%

Percentage of runs for which model was selected				
Pr(General)	**Pr(diff X and Y)**	**Pr (diff Y nondiff X)**	**Pr (nondiff Y diff X)**	**Pr(nondiff)**
25.5%	16.8%	23.4%	21.2%	13.0%

a. 500 simulation studies; 229 internal validation observations and 687 main study observations per simulation. b. True ln(OR)=1.14. c. α is the significance level used for LRT.

4.2.4 Study IV: Misclassification in Case-control studies

Table 4.9 summarizes simulations assessing our approach under the "case-control" sampling. We first generated 5000 cross-sectional observations, with a true OR of 2.7 (ln(OR)=1). Error-prone response Y^* and exposure X^* values were then generated with pre-specified misclassification probabilities. To mimic the case-oversampling, all the data with $Y^*=1$ (cases) were selected, while only 5% of observations with $Y^*=0$ (controls) were picked. This resulted in approximately 800 observations, from which ¼ of the total sample was then randomly selected as the internal validation subsample. When misclassification in Y is nondifferential, Table 4.9 suggests that the main/internal validation study-based analysis remained valid, regardless of the misclassification mechanism of X. However, the "operating" SE of Y is greater than the population SE_y, while the "operating" SP_y is smaller than the population SP_y. If misclassification on Y is nondifferential, however, the validity of the analysis fails to hold (results not shown).

Table 4.9 Results of Simulations Addressing Main/Internal Validation Study-Based Analysis Under Case-control Sampling[a]

Model	$\log(\widehat{OR})$ (SD)	95% CI Coverage
Nondifferential SE and SP Case: log(OR)=1; SE_x=0.7; SP_x=0.94;SE_y=0.85;SP_y=0.9		
Naive	0.14(0.16)	0
True	0.99(0.20)	95.8%
Main/Internal Validation	1.00(0.42)	96.0%
Nondifferential in Y and Differential in X: log(OR)=1; SE_{x1}=0.75; SP_{x1}=0.80; SE_{x0}=0.60; SP_{x0}=0.90;SE_y=0.60;SP_y=0.90		
Naive	0.17(0.16)	0
True	1.01(0.23)	97.0%
Main/Internal Validation	1.04(0.49)	96.7%

a. 500 simulation studies; roughly 200 internal validation observations and 600 main study observations per simulation based on 100% and 5% sampling of cases and controls, respectively.

4.3. EXAMPLE

The motivating example here is the HIV Epidemiology Research Study (HERS). This is a multi-center prospective cohort study with a total of 1310 women enrolled in four U.S. cities from 1993 to 1995 (43). Among them, 871 women were HIV-infected, and 439 were not infected but at risk. During each semi-annual visit, a large body of information was collected. The question of interest is to assess the association between the prevalence of bacterial vaginosis (BV) and the incidence of trichomoniasis. BV was measured by two different clinical methods: the clinically-based (CLIN) and the laboratory-based (LAB) methods. CLIN was a less accurate method that diagnoses BV

by evaluating multiple clinical criteria based on a modified Amsel's criteria (44), while LAB relies on a more sophisticated Gram-staining technique (45). The LAB method is more expensive and serves here as an arguable gold-standard, while the CLIN method is more cost-efficient and accessible. The presence of trichomoniasis was evaluated by a wet mount technique and a culture method. Wet mount is the clinical diagnostic tool that is estimated to have much lower sensitivity compared to culture testing for trichomoniasis (50). For both BV and trichomoniasis measurements, gold-standard and error-prone diagnoses are widely available for all patients in the HERS beyond visit 4, making HERS an excellent illustrative example to demonstrate the performance of the proposed validation data-based statistical methods.

We consider 916 patients with complete observations on both error-prone and gold-standard diagnoses of BV and trichomoniasis at the 4^{th} visit. The prevalence of BV via the LAB technique in the sample is around 18.2%, and after misclassifying the diagnoses, the naïve CLIN prevalence is about 7.5%. Compared to the LAB BV, the CLIN BV has a crude SE around 37% and SP about 99%, indicating that by using CLIN BV to assess the BV status, more than half of the BV positive patients are misdiagnosed as BV negative, while most BV-negative subjects are assessed correctly. The true prevalence of trichomoniasis in our sample is around 40.2% when assessed by culture testing. In contrast, when evaluated by wet mount, the prevalence is only 24.5%, with crude SE of 51.9% and SP of 94.0%. Like CLIN BV, wet mount diagnoses of trichomoniasis are relatively accurate when evaluating negative subjects, but the capability of capturing positive subjects is not satisfactory.

Table 4.10 Results of Analysis of 916 Women at Visit 4

Model	$\log(\widehat{OR})$ (SE)	$\widehat{OR}$ (95% CI)	P-value
Naive[a]	1.54(0.26)	4.65 (2.81, 7.69)	<0.0001
True[b]	1.14(0.18)	3.13 (2.21, 4.43)	<0.0001
Main/Internal Validation: Model 1[c]	1.18(0.33)	3.24 (1.14, 5.35)	0.0004
Main/Internal Validation: Model 2[d]	1.25(0.33)	3.48 (1.25, 5.71)	0.0001
Main/Internal Validation: Model 3[e]	1.50(0.32)	4.47 (1.63, 7.30)	<0.0001
Main/Internal Validation: Model 4[f]	1.34(0.32)	3.82 (1.39, 6.25)	0.0001
Main/Internal Validation: Model 5[g]	1.58(0.31)	4.84 (1.90, 7.78)	<0.0001

a.CLIN BV vs Wet Mount Trichomoniasis. b. LAB BV vs Culture Trichomoniasis c. 229 internal validation observations and 687 main study observations per simulation. Model 1 assuming dependent and differential misclassification. d. Model 2 assuming independent and differential misclassification. e. Model 3 assuming differential misclassification for Y and nondifferentiality for X. f. Model 4 assuming nondifferential misclassification for Y and differentiality for X.g. Model 5 assuming completely nondifferential misclassification.

Table 4.10 summarizes the results of using gold-standard measurements, error-prone diagnoses and fitting correction models under various misclassification mechanisms. The naïve result characterizing the association between CLIN BV and wet mount-based trichomoniasis inflated estimated OR by nearly 50% relative to the LAB and culture-based analyses. With main/internal validation analysis based on Model 1 through Model 5, using a random subsample accounting for ¼ of the total sample size as the internal validation set, the corrected $\widehat{OR}$ is close to the gold-standard (LAB and culture-based) result, though with expected efficiency loss, when dependent and differential misclassification is allowed (model 1). If independence with differentiality (model 2) is assumed, the corrected $\widehat{OR}$ appears biased away from the null. When nondifferential misclassification models (either or X or Y or both) are adopted, the corrected $\widehat{OR}$ is similar to the naïve result.

With the proposed model selection approach (Section 4.1.7), we first compare Model 2 vs. Model 1. The LRT test statistic=6.7 with df=1, and the p-value is 0.009. Therefore, we keep Model 1 as our final model, suggesting that in ther HERS example, one ideally needs to model dependent misclassification that is differential with respect to both X and Y.

4.4 Discussion

In this chapter, we have considered the classic problem of analyzing 2×2 tables, when both binary response and exposure variables are subject to misclassification. We place a heavy emphasis on specifying likelihood functions corresponding to main/internal validation designs, by expanding prior well-known matrix (1) and inverse matrix (8)

methods to a more general context. We also expand the idea of correcting the OR estimate from equation-based approaches to a more formal parametric approach, yielding a reliable estimate. Though matrix and inverse matrix methods can be viewed as special cases of the proposed approach, it distinguishes from prior efforts mainly in the flexibility of modeling more complex misclassification mechanisms, for instance, dependent and differential misclassification. This advantage may help to resolve many practical issues arising from data in this context, especially when the issues are seldom covered in the literature. It should be noted that matrix and inverse matrix methods are only equivalent to special cases of the proposed likelihood-based approach, when MLEs of misclassification rates are supplied into the generalized matrix identities given in this chapter. Otherwise, matrix and inverse matrix methods are not fully efficient. The likelihood can either be numerically optimized, the MLEs can be obtained explicitly when specified in its most general form based on the predictive value parameterization given in Section 4.1.3. If one is also interested in obtaining a confidence interval for the OR, the numerical optimization is recommended for its ease in computing standard errors.

We have proposed a straightforward model selection procedure for practitioners who not only seek to obtain a valid analytic result but also pursue a more precise result. It has been demonstrated that the proposed model selection procedure works stably in ensuring the correctly-specified model is often picked, while necessary model reduction is obtained. However, since the saturated model allowing dependent and differential misclassification is valid all the time, and appear to sacrifice relatively little efficiency given an adequate validation sample, it may often be prudent to recommend the choice of the saturated misclassification model (model 1).

We have also examined the impact of "case" oversampling on misclassification correction, which is a practical view of a case-control setting. in our context We have found that cautions need to be taken when analyzing such case-control data with misclassification. Only under certain situations, the proposed approach can be applied directly, yielding valid results. More specifically, with oversampling on Y^*, the approach holds if Y is nondifferentially misclassified, regardless of how X is misclassified. Similarly, with oversampling on X^*, how X is misclassified is crucial. If X is subject to nondifferential misclassification, the proposed approach works, no matter what misclassification process applies to Y. This result is consistent with prior findings (16).

The performance of the approach is demonstrated via a detailed analysis of the HERS example along with extensive simulation studies. It is important to note that when misclassification is differential, the resulting naïve ln(OR) can be biased in either direction; thus, applying a misclassification model that is sufficiently general is critical to getting a sensible result. More interestingly, when nondifferentiality does not hold, the corrected result based on that assumption may not even outdo the naïve result, as shown in the HERS example. For this reason, we urge readers not to simply assume nondifferentiality when analyzing data, unless the assumption is supported by the data or there are no other choices.

Future work could involve more consideration of cost-efficient designs of the internal validation sampling. Throughout the chapter, for convenience, we assume validation on both X and Y simultaneously in the internal validation sample. However, in practice, the costs associated with validating X or Y can be very different. It would be of interest to be able to allocate the validated observations cleverly into different types, to ensure the

control of the cost while still maintaining analytic validity, which is an extension of prior work (18). In some situations, the investigator may be more interested in validating a particular subpopulation, leading to nonrandom validation sampling. For example, one may care more about validating cases than controls. (52). There could also be interest in correction approaches when there is no gold standard available but one has access to replicates or an alloyed gold standard (53, 54).

Chapter 5 Misclassification in Response and Predictor Variables in Logistic Regression

5.1. Methods

5.1.1 Notation

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

$$logit[\Pr(Y = 1|X, C_1, C_2 \dots C_P)] = \beta_0 + \sum_{p=1}^{P} \beta_p c_p + \beta_{P+1} x \qquad (5.1),$$

where Y is the binary response variable, C_p (p=1,…P) denotes the pth covariate measured perfectly, and X stands for a binary predictor that is subject to misclassification. 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 the misclassification mechanism in the study (assumed in Chapter 4), a random sample of size n_v is selected, and gold standard measures of the response and exposure Y and X are made. Thus, the sample size of the main study will be n_m=n-n_v. If replacing X and Y in eqn. (5.1) with naïve measures X^* and Y^*, estimates of (β_0, …, β_p, β_{p+1}) can be potentially biased, and the magnitudes of biases rely on diagnostic properties of the methods used to classify X^* and Y^*.

5.1.2 Independent Nondifferential Misclassification

Assuming nondifferentiality, the misclassification parameters, known as sensitivity (SE) and specificity (SP), are constants that do not vary upon other information. For

example, regarding diagnostic properties relating Y^* to Y, in the nondifferential case, we define

$$SE_y = \Pr(Y^* = 1|Y = 1) \text{ and } SP_y = \Pr(Y^* = 0|Y = 0) \quad (5.2)$$

Similarly, to characterizing the method classifying X^*, we denote that

$$SE_x = \Pr(X^* = 1|X = 1) \text{ and } SP_x = \Pr(X^* = 0|X = 0) \quad (5.3).$$

Note that SE and SP presented here are constants, and they are independent of other information, such as disease status and prognostic factors. In other words, we assume that $\Pr(Y^* = 1|Y = 1) = \Pr(Y^* = 1|Y = 1,\ x, \boldsymbol{c})$.

For the simplicity of illustration, we first consider the situation when both X and Y are subject to nondifferential misclassification. Following the rule of total probability, each independent observation in the main study contributes to the following likelihood term:

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

$$= \sum_{y=0}^{y=1}\sum_{x=0}^{x=1} Pr(y^*|y)Pr\,(x^*|x)Pr(y|x, \boldsymbol{c})Pr(x|\boldsymbol{c}) \quad (5.4)$$

The first and second terms in eqn. (5.4) represent the SE/SP of Y and X, while the third term reflects the primary model of interest that is defined in eqn. (5.1). The last term characterizes the association of X with other covariates $\boldsymbol{C}$. Note that in fact the vector $\boldsymbol{C}$ in the latter model may or may not be exactly the same as the vector $\boldsymbol{C}$ in the primary model, though throughout this chapter we do not rotationally distinguish the two. To facilitate the likelihood representation, a model for $\Pr(X|\boldsymbol{C})$ needs to be specified. Here

we adopt the familiar logit link as in eqn. (5.5), as X is a binary variable. However, other links can be applied without conceptual difficulty.

$$logit\left[\Pr\left(X = 1 \middle| C_1, C_2 \ldots C_Q\right)\right] = \gamma_0 + \sum_{q=1}^{Q} \gamma_q \, c_q \quad (5.5)$$

Assuming nondifferential misclassification in both X and Y, we can repeat the likelihood of the main study as follows:

$$L_m = \prod_{i=1}^{n_m} \left\{ \sum_{y_i=0}^{y_i=1} \sum_{x_i=0}^{x_i=1} Pr\left(y^*_i \middle| y_i\right) Pr\left(x^*_i \middle| x_i\right) Pr\left(y_i \middle| x_i, \boldsymbol{c_i}\right) Pr\left(x_i \middle| \boldsymbol{c_i}\right) \right\} \qquad (5.6).$$

The estimates for SE_x, SE_y, SP_x and SP_y are assumed to come from extra data, such as internal validation data which is the focus of this chapter. In the internal validation subsample, we assume that (X, X^*,Y,Y^*) are observed on each subject. The likelihood contribution from the subsample is:

$$L_v = \prod_{j=1}^{n_v} Pr\left(y^*_i \middle| y_i\right) Pr\left(x^*_i \middle| x_i\right) Pr\left(y_i \middle| x_i, \boldsymbol{c_i}\right) Pr\left(x_i \middle| \boldsymbol{c_i}\right) \quad (5.7)$$

5.1.3 Independent Differential Misclassification

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

$$SE_{yec} = \Pr(Y^* = 1 | Y = 1, X = e, \boldsymbol{C} = \boldsymbol{c}) \text{ and}$$

$$SP_{yec} = \Pr(Y^* = 0 | Y = 0. X = e\,, \boldsymbol{C} = \boldsymbol{c}) \qquad (5.8).$$

As opposed to nondifferential misclassification as described in Section 5.2.2, the SE and SP of Y now can be functions of exposure (X) and other covariates ($\boldsymbol{C}$). For clarification, as long as SE and SP depend only on the true values of other variables, we call it "differential but independent misclassification". Eqn. (5.8) is the most general representation of SE_y and SP_y under such an assumption. However, the definition of $\boldsymbol{C}$ is flexible upon model selection, and it may not be the same as the $\boldsymbol{C}$ in the primary model eqn. (5.1). It may share overlap with the covariate vector in the primary model, or may include factors not included in eqn. (5.1). In the meantime, X may or may not be an important factor in characterizing the diagnostic properties for Y, and it can be left out when it is deemed not to be. Similarly, we may define the misclassification process of X as:

$$SE_{xdc} = \Pr(X^* = 1 | X = 1, Y = d, \boldsymbol{C} = \boldsymbol{c}) \text{ and}$$

$$SP_{xdc} = \Pr(X^* = 0 | X = 0. Y = d, \boldsymbol{C} = \boldsymbol{c}) \quad (5.9).$$

Again, like in Section 5.1.2, each observation in the main study contributes to the likelihood as follows:

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

$$= \sum_{\mathrm{y}=0}^{\mathrm{y}=1} \sum_{\mathrm{x}=0}^{\mathrm{x}=1} \Pr(\mathrm{y}^* | \mathrm{y}, \mathrm{x}, \boldsymbol{c}) \Pr(\mathrm{x}^* | \mathrm{x}, \mathrm{y}, \boldsymbol{c}) \Pr(\mathrm{y} | \mathrm{x}, \boldsymbol{c}) \Pr(\mathrm{x} | \boldsymbol{c}) \quad (5.10)$$

The last two terms of eqn. (5.10) are described in Section 5.1.2. The first two terms, characterizing SE and SP for X and Y, need to be fully 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, we define

$$logit[\Pr(Y^* = 1|Y, X, C_1, C_2 \dots C_R)] = \theta_0 + \sum_{r=1}^{R} \theta_r c_r + \theta_{R+1} y + \theta_{R+2} x \quad (5.11)$$ and

$$logit[\Pr(X^* = 1|Y, X, C_1, C_2 \dots C_S)] = \delta_0 + \sum_{s=1}^{S} \delta_r c_r + \delta_{S+1} x + \delta_{S+2} y \quad (5.12)$$

Eqn. (5.11) implies that

$$SE_{yec} = \frac{\exp(\theta_0 + \sum_{r=1}^{R} \theta_r c_r + \theta_{R+1} + \theta_{R+2} x)}{1 + \exp(\theta_0 + \sum_{r=1}^{R} \theta_r c_r + \theta_{R+1} + \theta_{R+2} x)}$$

and

$$SP_{yec} = 1 - \frac{1}{1 + \exp(\theta_0 + \sum_{r=1}^{R} \theta_r c_r + \theta_{R+2} x)}$$

The corresponding definition of SE_{xdc} and SP_{xdc} follows simultaneously from model (5.12). The full likelihood is $L=L_m x L_v$, where

$$L_m = \prod_{i=1}^{n_m} \left\{ \sum_{y_i=0}^{y_i=1} \sum_{x_i=0}^{x_i=1} Pr(y^*_i | y_i, x_i, \boldsymbol{c_i}) Pr(x^*_i | x_i, y_i, \boldsymbol{c_i}) Pr(y_i | x_i, \boldsymbol{c_i}) Pr(x_i | \boldsymbol{c_i}) \right\} \quad (5.13)$$

and

$$L_v = \prod_{j=1}^{n_v} Pr\left(y^*_j \middle| y_j, x_j, \boldsymbol{c_j}\right) Pr\left(x^*_j \middle| x_j, y_j, \boldsymbol{c_j}\right) Pr\left(y_j | x_j, \boldsymbol{c_j}\right) Pr\left(x_j \middle| \boldsymbol{c_j}\right) \quad (5.14)$$

The major distinction of eqn.s (5.13) and (5.14) from eqn. (5.6) and (5.7) is that the likelihood representing differential misclassification incorporates the modeling of the SE

and SP of X and Y based on (5.11) and (5.12). Nondifferential misclassification may be viewed as a special case of differential misclassification, since it indicates that (θ_1=…=θ_R=θ_{R+2}=0) or (δ_1=…=δ_S=δ_{S+2}=0). The likelihood-based approach allows hypothesis testing to assess the null hypothesis of nondifferentiality, as well as model selection for screening out factors associated with SE and SP for both X and Y.

5.1.4 Dependent and Differential Misclassification

In Section 5.1.3, SE and SP of X and Y were allowed to be impacted by the true values of other factors, noted as "differential misclassification". Another type of misclassification what we will refer to as "dependent misclassification", when SE and SP depend on error-prone values. The likelihood for the saturated model (shown in eqn. (5.15)) is an example of the combination of dependence and differentiality. In particular, eqn. (5.15) implies that $\Pr(y^*,x^*,y,x|c) \neq \Pr(y^*,y,x|c)\Pr(x^*,y,x|c)$, where $\Pr(y^*,x^*,y,x|c) = \Pr(y^*,y,x|c)\Pr(x^*,y,x|c)$ suggests statistical independence. So more generally,

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

$$= \sum_{y=0}^{y=1}\sum_{x=0}^{x=1} Pr(y^*|y, x, x^*, \boldsymbol{c}) Pr\,(x^*|x, y, \boldsymbol{c}) Pr(y|x, \boldsymbol{c}) Pr(x|\boldsymbol{c}) \quad (5.15)$$

Note that eqn. (5.15) is the example of a saturated model with both dependent and differential misclassification present; the dependence is implied by the conditioning or X^* in the firs tterm. The full likelihood is still $L=L_m x L_v$, where

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

and

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

With proper model selection, it is possible to reduce the saturated model to a simpler form. The hypothesis testing for independence for eqn. (5.15) is to test whether the logistic regression coefficient for x^* (θ_{R+3} in eqn. (5.16)) from the SE/SP model of Y is zero or not.

$$logit[\Pr(Y^* = 1 | Y, X, C_1, C_2 \dots C_r)] = \theta_0 + \sum_{i=1}^{R} \theta_r c_r + \theta_{R+1} y + \theta_{R+2} x + \theta_{R+3} x^* \quad (5.16)$$

5.2.5 Other Types of Misclassification

It should be noted that besides the misclassification mechanisms displayed in Sections 5.2.2-5.2.4, there are other possible types of mechanisms. For example, in practice, it is possible that both exposure and disease status are nondifferentially misclassified, but misclassification is dependent (41). When this is the case, we may adjust the form of the likelihood as follows:

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

and

$$L_v = \prod_{j=1}^{n_v} Pr\left(y^*_i \middle| y_i, x^*_i\right) Pr(x^*_i|x_i)\ Pr\left(y_i|x_i, \boldsymbol{c_i}\right) Pr(x_i|\boldsymbol{c_i})$$

Other types may include the possibility that only disease or exposure status is independently nondifferentially misclassified. For example, if only disease status is independently nondifferentially misclassified, we have the likelihood as:

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

and

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

The likelihoods presented in Section 5.2 can be numerically optimized via standard statistical software such as SAS NLMIXED (57). In practical analysis, thorough model selection should be performed for the SE/SP models for X and Y using the validation sample, to assess whether dependence and/or differentiality is involved in the misclassification process.

5.2. Example

We consider data on bacterial vaginosis (BV) and trichomoniasis (Trich) status for women in the HIV Epidemiology Research Study (HERS) as an illustrative example.

The HERS data at the 4th visit is used with 904 women with complete data on BV, TRICH and other risk factors. The medium age at enrollment was 37. Among them, 61.7% of women were blacks, 67.4% were HIV positive and 52% were intravenous drug users.

Using culture testing, 18% of women were diagnosed with trichomoniasis. Since wet mount is not sensitive, only 7.6% were trichomoniasis-positive based on wet mounts. 40.3% of women were BV positive via the LAB method, compared to only 24.5% based on the CLIN method. The crude sensitivities for the wet mount and CLIN methods were only 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. In the HERS data, (X,X^*,Y,Y^*) were measured on all participants, providing an ideal data example to illustrate the performance of the proposed approach.

We first fit eqn. (5.1) on all subjects by using Y and X as the response and predictor variables, referred to as "Ideal Analysis". Preliminary model selection suggests that trichomoniasis status, age, race, HIV risk cohort (RISKCHRT) and HIV status (HIVPOS) are important risk factors for BV, as shown in eqn. (5.17).

$$logit\left[\Pr(LAB\ BV = 1)\right] = \beta_0 + \beta_1 CULTURETRICH + \beta_2 AGE + \beta_3 RACE + \beta_4 RISKCHRT + \beta_5 HIVPOS \quad (5.17)$$

We then fit the same model by replacing X and Y with X^* and Y^*, denoted as "Naïve Analysis". The results are summarized in Table 5.1. The two analyses differ markedly in magnitudes of the estimated OR for trichomoniasis (2.41 for ideal vs. 3.44 for naïve) and HIV risk cohort (1.37 for ideal vs. 2.45 for naïve). The estimated ORs for HIV status differ in directionality (1.25 for ideal vs. 0.73 for naïve).

Table 5.1 Logistic Regression Results on 904 Women at 4th Visit.

Variable	$\widehat{\beta}$ (StdErr)	$\widehat{OR}$ (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.CLIN BV vs Wet Mount Trichomoniasis, adjusting for age, race, HIV risk cohort and HIV staus. b. LAB BV vs Culture Trichomoniasis , adjusting for age, race, HIV risk cohort and HIV staus.

In order to demonstrate the performance of the proposed approach, we randomly selected 1/3 of the total sample size (n_v=214) into the internal validation subsample. Model selection on those 214 women suggested a version of the X|C model as follows:

$$logit\,[\Pr(CULTURE\ TRICH = 1)] = \gamma_0 + \gamma_1 AGE + \gamma_2 RACE + \gamma_3 RISKCHRT \quad (5.18)$$

where age, race and HIV risk cohort are associated with trichomoniasis status. Predictor selection appiled to these 214 women future suggested dependent and differential misclassification in the CLIN BV and WET TRICH methods. The selected SE/SP models for CLIN BV and WET TRICH are as shown in eqns. (5.19-5.20).

$$logit[\Pr(CLINBV = 1)] = \theta_0 + \theta_1 LABBV + \theta_2 WETTRICH + \theta_3 RISKCHRT + \theta_4 HIVPOS \quad (5.19)$$

$$logit[\Pr(WETTRICH = 1)] = \delta_0 + \delta_1 CULTURETRICH + \delta_2 RISKCHRT \quad (5.20)$$

More specifically, the classification rates of CLIN BV are dependent on risk factors HIV risk cohort and HIV status, implying differential misclassification. In the meanwhile, the misclassification process for BV also depends on the error-prone wet mount version of the trichomoniasis diagnosis, implying the presence of dependent misclassification. Similarly, HIV risk cohort has a significant impact on the SE and SP of the wet mount method, a typical example of differential misclassification.

The first model fitted in Table 5.2 is the complete analysis of the data by jointly modeling eqn.s (5.17)-(5.20), yielding the same interpretation as the ideal analysis, with all primary parameter estimates having similar magnitudes and the same directionalities. In contrast, results assuming independent differential or nondifferential misclassification are more similar to the results of the naïve analysis. For instance, when assuming independence but allowing differential misclassification, the estimate for trichomoniasis status largely increases, with a similar magnitude as observed in the naïve analysis. If assuming independent and nondifferential misclassification, a greatly elevated estimate for HIV risk cohort and a negative estimated ln(OR) for HIV status are also noticed, besides the bias in trichomoniasis. Unsurprisingly, the likelihood ratio tests comparing the two simpler models in Table 5.2 with the most general misclassification model highly significant, strongly suggesting a need to account for dependent and differential misclassification (χ^2=15.4, p<0.0001 for comparing independent differential model with dependent and differential model; χ^2=40.4, p<0.0001 for comparing independent nondifferential model with dependent and differential model). This clearly highlights the

importance of internal validation sampling for evaluating and modeling complicated misclassification mechanisms.

Tables 5.3-5.4 summarize the maximum likelihood estimates of SE and SP corresponding to CLIN BV and wet mount trichomoniasis in different strata. For CLIN BV, wet mount trichomoniasis diagnoses, 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 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 seem to have a greater SE than those at risk via sexual contact, while the test is similarly highly specific in both groups.

Table 5.2 Results of Maximum Likelihood Analysis of Main/Internal Validation Study Data on 904 Women at 4th Visit (n_m=690, n_v=214).

Variable	$\widehat{\beta}$ (StdErr)	$\widehat{OR}$ (95% CI)
Assuming dependent and differential misclassification		
Trichomoniasis (+ vs. -)	0.76 (0.40)	2.13 (0.44, 3.82)
Age (Years)	-0.05 (0.02)	0.95 (0.92, 0.98)
Race (Black vs. Others)	0.80 (0.23)	2.22 (1.19, 3.26)
HIV Risk Cohort (IDU vs. Sexual)	0.28 (0.26)	1.33 (0.65, 2.01)
HIV Status (+ vs. -)	0.22 (0.27)	1.24 (0.58, 1.91)
Assuming independent and differential misclassification		
Trichomoniasis (+ vs. -)	1.33 (0.37)	3.78 (1.02, 6.54)
Age (Years)	-0.05 (0.02)	0.95 (0.92, 0.98)
Race (Black vs. Others)	0.76 (0.24)	2.13 (1.14, 3.11)
HIV Risk Cohort (IDU vs. Sexual)	0.32 (0.26)	1.38 (0.68, 2.09)
HIV Status (+ vs. -)	0.16 (0.27)	1.17 (0.55, 1.80)
Assuming nondifferential misclassification		
Trichomoniasis (+ vs. -)	1.51 (0.38)	4.51 (1.17, 7.86)
Age (Years)	-0.05 (0.02)	0.96 (0.92, 0.99)
Race (Black vs. Others)	0.71 (0.24)	2.04 (1.09, 2.98)
HIV Risk Cohort (IDU vs. Sexual)	0.81 (0.22)	2.25 (1.26, 3.25)
HIV Status (+ vs. -)	-0.15 (0.23)	0.86 (0.47, 1.25)

a. Maximum likelihood estimates of primary parameters are obtained by jointly modeling eqn.s (5.17)-(5.20). b. WETTRICH is removed from eqn. (5.19) to indicate independence. The assumption is not supported by the data ($p<0.0001$). c. No covariates affect SE and SP of Y and X in eqn.s (5.19) and (5.20). The assumption is not supported by the data ($p<.0001$).

Table 5.3 Results of Maximum Likelihood Analysis of Main/Internal Validation Study Data on 904 Women at 4th Visit (n_m=690, n_v=214): Estimates of SE and SP of CLIN BV[a].

Wet Mount Trichomoniasis	HIV Risk Cohort	HIV Status	$\widehat{SE}$ (StdErr)	$\widehat{SP}$ (StdErr)
+	IDU	+	0.83 (0.07)	0.81 (0.08)
+	IDU	-	0.90 (0.05)	0.69 (0.11)
+	Sexual	+	0.67 (0.11)	0.91 (0.05)
+	Sexual	-	0.79 (0.09)	0.85 (0.07)
-	IDU	+	0.51 (0.05)	0.95 (0.02)
-	IDU	-	0.66 (0.07)	0.92 (0.03)
-	Sexual	+	0.29 (0.04)	0.98 (0.01)
-	Sexual	-	0.44 (0.07)	0.96 (0.02)

a. Std errors obtained by multivariate delta method.

Table 5.4 Results of Maximum Likelihood Analysis of Main/Internal Validation Study Data on 904 Women at 4th Visit (n_m=690, n_v=214): Estimates of SE and SP of Wet Mount Trichomoniasis.[a]

HIV Risk Cohort	$\widehat{SE}$ (StdErr)	$\widehat{SP}$ (StdErr)
IDU	0.51 (0.08)	0.99 (0.01)
Sexual	0.23 (0.06)	0.99 (0.003)

a. Std errors obtained by multivariate delta method.

5.3. Simulation Studies

The simulation experiment summarized in Table 5.5 demonstrates the performance of jointly modeling eqns. (5.17-20) under conditions similar to the HERS example described in Section 5.2. The predictor subject to misclassification (X) along with three covariates (C_1-C_3) were generated with distributions mimicking the observed data in the HERS study at visit 4, i.e., mimicking trichomoniasis, age, HIV risk cohort and HIV status. The true response Y was simulated under eqn. (5.17). Error prone outcome and predictor (Y^* and X^*) were generated via eqns. (5.19-5.20), with true coefficients similar to these estimated in the ideal analysis in Table 5.1. With 500 simulated datasets, ideal, naïve and complete analyses were conducted on each dataset. Table 5.5 suggests that the naïve analysis yields greatly biased estimates. Assuming dependent and differential misclassification, note that a complete analysis produces reliable results and excellent 95% confidence interval coverage.

Table 5.5 Results of Simulations Designed to Mimic Conditions of HERS Example[a].

Variable	$\widehat{\boldsymbol{\beta}}$ (StdErr)	95% CI Coverage
Ideal Analysis[b]		
Trichomoniasis (+ vs. -)	0.89 (0.27)	94.8%
Age (Years)	-0.04 (0.003)	93.2%
Race (Black vs. Others)	0.77 (0.18)	96.0%
HIV Risk Cohort (IDU vs. Sexual)	0.31 (0.19)	94.8%
HIV Status (+ vs. -)	0.22 (0.19)	94.2%
Naïve Analysis[c]		
Trichomoniasis (+ vs. -)	1.35 (0.28)	63.2%
Age (Years)	-0.02 (0.002)	0
Race (Black vs. Others)	0.30 (0.17)	25.0%
HIV Risk Cohort (IDU vs. Sexual)	0.89 (0.19)	10.4%
HIV Status (+ vs. -)	-0.42 (0.18)	4.8%
Complete Analysis[d]		
Trichomoniasis (+ vs. -)	0.91 (0.52)	94.2%
Age (Years)	-0.04 (0.006)	94.8%
Race (Black vs. Others)	0.79 (0.30)	96.4%
HIV Risk Cohort (IDU vs. Sexual)	0.35 (0.34)	94.8%
HIV Status (+ vs. -)	0.22 (0.34)	94.4%

a. 500 simulations. n_m=690, n_v=214. Maximum likelihood estimates of primary parameters are obtained by jointly modeling eqn.s (5.17)-(5.20). True parameters: (β_0=0.14, β_1=0.88, β_2=-0.04, β_3=0.76, β_4=0.31, β_5=0.22). ($\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$) are set to equal MLEs from HERS analysis (not shown). b. MLEs from eqn. (5.17). c. MLEs from eqn. (5.17) with (Y^*, X^*) replacing (Y, X). c. Maximum likelihood estimates of primary parameters are obtained by jointly modeling eqn.s (5.17)-(5.20).

5.4. Discussion

In this chapter, we have expanded the ML approach proposed in Chapter 4 to a general regression setting, by following parametric ideas outlined in Lyles et al (16). Our goal is to provide clear guidance on adjusting for biases due to misclassification in binary response and predictor variables in ordinary logistic regressions. We strongly emphasize the importance of using internal validation sampling to assess misclassification patterns and to ensure the validity of the results. The approach outlined in this chapter provides a general and reliable way to flexibly model a wide variety of misclassification mechanisms, including dependent and differential misclassification. The parametric model makes likelihood ratio testing an option to assist with model selection and mechanism evaluation, as demonstrated by the HERS example. Though throughout logit links are adopted for all models, other links can also be used without additional conceptual and technical difficulty.

As with any parametric model, to correctly specify the model is crucial in terms of obtaining valid estimates. However, empirical evidence suggests that the approach enjoys some robustness to misspecification of the X|C and SE/SP models (not shown), requiring further investigation.

There may be interest in extending the regression-based correction approach to adjust for outcome and predictor misclassification in situations when both are repeatedly measured, representing a longitudinal setting. To address this question, methods described in Chapter 2 and 3 may aid in the methodology development.

REFERENCES

1. Barron, B.A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics*. **33**, 414-7.

2. Copeland, K.T., Checkoway, H., McMichael, A.J. & Holbrook, R.H. (1997). Bias due to misclassification in the estimation of relative risk. *Am. J. Epidem*. **105**, 488-95.

3. Carroll R.J., Ruppert D, Stefanski L.A., Crainiceanu C.M. Measurement Error in Nonlinear Models, Second Edition. London: Chapman and Hall, 2006.

4. Neuhaus J.M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*. **86**, 843-55.

5. Green, M.S. (1983). Use of predictive value to adjust relative risk estimates biases by misclassification of outcome status. *Am.J.Epidem*. **117**, 98-105.

6. Greenland, S. (1988). Variance estimation of epidemiologic effect estimates under misclassification. *Statist. Med*. **7**, 745-57.

7. Brenner, H. & Gefeller, O. (1993). Use of positive predictive value to correct for diease misclassification in epidemiologic studies. *Am. J. Epidem*. **138**, 1007-15.

8. Marshall R.J. (1990). Validation study methods for estimating proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*. **43**, 941-947.

9. Morrissey M.J., Spiegelman D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics*. **55**, 338-344.

10. Greenland S. (2008). Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. *Journal of Statistical Planning and Inference*. **138**, 528-538.

11. Magder L.S., Hughes J.P. (1997). Logistic regression when the outcome is measured with uncertainty..*American Journal of Epidemiology*. **146**, 195-203.

12. Paulino C.D., Soares P, Neuhaus J. (2003). Binomial regression with misclassification. *Biometrics*. **59**, 670-75.

13. McInturff P, Johnson W.O., Cowling D., Gardner I.A. (2004). Modeling risk when binary outcomes are subject to error. *Statistics in Medicine*. **23**, 1095-1109.

14. Gerlach R, Stamey J. (2007). Bayesian model selection for logistic regression with misclassified outcomes. *Statistical Modelling*. **7**, 255-73.

15. Pepe M.S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika*. **79**, 355-65.

16. Lyles, R.H., Tang, L., Superak, H.M., King, C.C., Celantano, D., Lo, Y., Sobel, J. (2011). An illustration of validation data-based adjustments for outcome misclassification in logistic regression. *Epidemiology*. **22**, 589-97.

17. Neuhaus, J.M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*. **58**,675-73.

18. Lyles R.H., Williamson, J.M., Lin, H.M., Heiling C.M. (2005). Extending McNemar's test: estimation and inference when paired binary outcome data are misclassified. *Biometrics*. **61**, 281-94.

19. Bross I.D.J. (1954). Misclassification in 2×2 tables. *Biometrics*. **10**:478-486.

20. Copeland K.T., Checkoway H., McMichael,A.J. and Holbrook, R. H. (1977). Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology*. **105**, 488-95.

21. Walker A.M. and Blettner M. Comparing imperfect measures of exposure. (1985). *American Journal of Epidemiology*. **121**, 783-790.

22. Dosemeci M., Wacholder S. and Lubin J. H. (1990). Does nondifferential misclassification of exposure always bias a true effect toward the null value? *American Journal of Epidemiology*. **132**, 746-748.

23. Chavance M., Dellatolas G. and Lellouch J. (1992). Correlated nondifferential misclassification of disease and exposure. *International Journal of Epidemiology*. **21**, 537-546.

24. Kristensen P. (1992). Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology*. **3**, 210-215.

25. Greenland S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*. **112**, 564-569

26. Greenland S. and Kleinbaum D. G. (1983). Correcting for misclassification in two way tables and matched-pair studies. *International Journal of Epidemiology*. **12**, 93-97.

27. Robins J.M., Rotnitzky A. and Zhao L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. **89**, 846-866.

28. Hui S.L. and Walter S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*. **36**, 167-171.

29. Marshall J. R. and Graham S. (1984). Use of dual responses to increase the validity of case-control studies. *Journal of Chronic Diseases*. **37**, 125-136.

30. Walter S.D. (1984). Use of dual responses to increase the validity of case-control studies: a commentary. *Journal of Chronic Diseases*. **37**, 137-139.

31. Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*. **41**, 959-968.

32. Rindskopf D. and Rindskopf W. (1986). The value of latent class analysis in medical diagnosis. S*tatistics in Medicine*. **5**, 21-27.

33. Liu X. and Liang K.Y. (1991). Adjustment for non-differential misclassification error in the generalized linear model. *Statistics in Medicine*. **10**, 1197-1211

34. Brenner H. (1992). Use and limitations of dual measurements in correcting for non-differential exposure misclassification. *Epidemiology*. **3**, 216-222.

35. Drews C.D., Flanders W.D. and Kosinski, A.S. (1993). Use of two data sources to estimate odds ratios in case-control studies. *Epidemiology*. **4**, 327-335.

36. Fox M.P., Lash T.L., and Geenland S. (2005). A method to automate probabilistic sensitivity analyses of misclassified binary variables. *International Journal of Epidemiology*. **34**, 1370-1376.

37. Gustafson P., Le N.D., Saskin R. (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics*. **57**, 598-609.

38. Lyles R.H. and Lin J. (2010). Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Statistics in Medicine*. **29**, 2297-2309.

39. Kosinski A.S. and Flanders W.D. (1999). Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification; a regression approach. *Statistics in Medicine*. **18**, 2795-2808.

40. Kleinbaum, D., Kupper, L., and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, California: Lifetime Learning.

41. Rothman K.J., Greenland S., Lash T.L. Modern Epidemiology, Third Edition. Lippincott Williams & Wilkins, 2008.

42. Lyles, R.H. (2002). A note on estimating crude odds ratios in case–control studies with differentially misclassified exposure. *Biometrics.* **58**, 1034–1037.

43. Smith D.K., Warren D.L., Vlahov D., Schuman P., Stein M.D., Greenberg B.L.(1997). 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.* **146**:459-469.

44. Amsel R., Totten P.A., Spiegel C.A., Chen K.C., Eschenbach D., Holmes K.K. (1983). Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. *American Journal of Medicine.* **74**:14-22.

45. Nugent R.P., Krohn M.A., Hillier S.L.(1991). Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *Journal of Clinical Microbiology.* **29**:297-301.

46. Breslow N.E. and Clayton D. G. (1993). Approximate inference in generalized linear mixed models. *JASA*. **88**, 9-25.

47. Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model. *Journal of Computational and Graphical Statistics*. **4**,12-35.

48. Beal, S.L. and Sheiner, L.B. (1988). Heteroskedastic Nonlinear Regression. *Technometrics*. **30**, 327-338.

49. SAS Institute, Inc. (2004). SAS/STAT 9.1 User's Guide. Cary, NC: SAS Institute, Inc.

50. Thomason J.L., Gelbart S.M., Sobun J.F., Schulien M.B. and Hamilton P.R. (1988). Comparison of four methods to detect Trichomonas vaginalis. *J Clin Microbiol*. **26**, 1869-1870.

51. Demirezen S., Korkmaz E. and Beksac M.S. (2005). Association between trichomoniasis and bacterial vaginosis: examination of 600 cervicovaginal smears. *Cent Eur J Public Health*. **13**, 96-98.

52. Spiegelman, D., Carroll, R., Kipnis, V. (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statistics in Medicine*. **10**, 139–160.

53. Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. (1993). *Am J Epidemiol*. **137**:1251–1258.

54. Brenner H. Correcting for exposure misclassification using an alloyed gold standard. (1996). *Epidemiology*. **7**:406–410.

55. Huang, Y. and Wang, C. Y. (1999). Nonparametric correction to errors in covariates. Technical Report, Fred Hutchinson Cancer Research Center, Seattle.