Outcome Misclassification in Logistic Regression: Examining Hospitalization Risk and its Association with Health Literacy

By

Brooklyn Stanley

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

March 31, 2019

Nashville, Tennessee

Approved:

Jonathan S. Schildcrout, Ph.D.

Robert A. Greevy, Ph.D.

Table of Contents

List of Figures

List of Tables

1 Introduction

In any statistical examination where regression techniques are utilized, the objective is to acquire a scientific model that portrays the connection between observations of the outcome or dependent variable, to be signified as T, and a collection of autonomous factors, referred to in total, without reference to their number and estimation scale, as X. A straightforward approach to conceptualize the model is to think about the estimate of the dependent variable as consisting of two components: the systematic component and the error component. The systematic component is the mathematical function of independent factors that portrays the "standard" estimation of T. The error component quantifies how much an individual subject's value of T varies from what is to be expected given their values of X. Common reasons for this error in the estimation model are the exclusion of important independent factors, incorrect model specification, or inherent error in the data collection, such as misclassification.

In general regression problems, covariates and outcomes are regularly measured with random error; in the case of discrete factors, random error is referred to as misclassification. While stochastic error models have received much consideration in writing, they are not easily extrapolated to the case of misclassification, as the distribution of the error is not continuous. It has been shown that disregarding possible misclassifications in analyses can lead to significant biases in the coefficient estimates and subsequent inferences (Magder & Hughes, 1997). Without careful consideration, misclassification in the outcome can be overlooked in the model, thus causing misleading results.

There are many cases where data collection is prone to misclassification. Often, misclassification is due to a systemic error in the test for the outcome. This error in the testing mechanism can be well-defined through replication and classified through two quantities: sensitivity and specificity. Methods exist to adjust for the bias induced by misclassification in these cases where the sensitivity and specificity are known. In many practical situations, however, the exact values of these error rates are unknown and difficult, if not impossible, to obtain. A clear-cut solution in these cases may not be possible, but the bias in the estimates should be addressed. Challenges arise because the proportion of misclassified outcomes will almost always be unknown; moreover, the misclassification could be dependent on other variables in the study. Instances like this are not uncommon, making methods to adjust for the biases in the estimates necessary for the validity of the results.

To demonstrate the problems with ignoring potential biases caused by misclassified outcomes, I evaluate data from the Mid-South Coronary Heart Disease Cohort Study (MCHDCS), which was developed as one of three populations of study under the Patient-Centered Outcomes Research Institute (PCORI) funded by Mid-South Clinical Data Research Network (CDRN), where survey data was collected from patients visiting a primary care facility. The patients were followed for one year post study enrollment to evaluate whether they were hospitalized during this time. The goal of the study was to evaluate the association between health literacy and hospitalization. The original data collection included information from hospitalizations that occurred only at Vanderbilt Medical Center. However, since Vanderbilt Medical Center is a tertiary care facility, it is likely that patients could have been hospitalized elsewhere in the area if their injuries or ailments were not severe. To account for this, the data were expanded to include hospitalizations in three additional hospitals in the Vanderbilt Health Affiliated Network (VHAN) - West TN Health, Maury Regional, and Williamson Medical Center. After collecting data from the surrounding hospitals, it was discovered that the exclusive use of Vanderbilt data would result in many misclassifications, as there was a substantial number of patients that were hospitalized elsewhere during the one-year follow-up. Furthermore, it is suspected that some of the variables studied could be associated with the hospital to which the patient was admitted. To illustrate the bias this invokes, I compare the results of the naive logistic regression with that of the true model and explore adjustment methods that have been proposed to adjust for misclassification.

2 Misclassification Adjustment Methods

Misclassification can be divided into two types: differential and non-differential. Differential misclassification occurs when the probability of being misclassified depends on all or a subset of the covariates, X. Non-differential misclassification occurs randomly, independent of the covariates in the model. This is an important distinction because the type of misclassification affects the direction of the biases: non-differential is typically

biased towards the null; however, differential misclassification can bias estimates in any direction, depending on the strength and direction of the variable-misclassification association. In either case, the error in the outcome will inevitably lead to misinterpretation of the estimated association between the variables and the outcome.

In cases of misclassification, we do not observe the true outcome, T, but rather an error-corrupted version, Y. The accuracy of Y can be quantified with the probabilities of observing the true value of T in Y: the sensitivity ($P(Y = 1|T = 1, X)$) and the specificity ($P(Y = 0|T = 0, X)$). For the purposes of most adjustment methods, the false positive probability, γ_0, and the false negative probability, γ_1, are typically derived from these quantities:

$$1 - spec = \gamma_0(X) = P(Y = 1|T = 0, X) \qquad 1 - sens = \gamma_1(X) = P(Y = 0|T = 1, X)$$

In the special case of non-differential misclassification, these quantities are independent of the variables in the model and thus fixed for all X:

$$\gamma_0(X) = P(Y = 1|T = 0) = \gamma_0 \qquad \gamma_1(X) = P(Y = 0|T = 1) = \gamma_1$$

For the example data, there was no misclassification observed when a patient was hospitalized at VUMC: all observed VUMC hospitalizations were still valid even after the addition of the VHAN data. So the specificity in the study was 1, yielding a false positive probability of $\gamma_0(X) = 0$. There is suspected differential misclassification in the example data, so the probability of a false negative, $\gamma_1(x)$, would not be assumed to be a constant.

Using these quantities, researchers have developed methods that utilize the relationship between specificity, sensitivity, and the outcome to adjust for the misclassification in the model and produce unbiased results. I will explore some of these methods and use the example hospitalization data to demonstrate their results.

2.1 Neuhaus Adjustment Method

John Neuhaus (Neuhaus 1999) was one of the first to explore potential adjustment methods for misclassification. He exploited the relationship between specificity, sensitivity, the naive outcome, and the true outcome to develop methods to estimate the bias in coefficient estimates. Using conditional probabilities, he derived the basic equation for the probability of observing $Y = 1$ in the naive outcome:

$$P_T(Y = 1|X) = \gamma_0(X)P(T = 0|X) + (1 - \gamma_1(X))P(T = 1|X) \tag{1}$$

Substituting the inverse link function, $g^{-1}(X\beta)$ for P(T = 1|X) in equation (1):

$$= \gamma_0(x)\big(1 - g^{-1}(\beta X)\big) + \big(1 - \gamma_1(X)\big)g^{-1}(\beta X) \tag{2}$$

$$= \big(1 - \gamma_0(X) - \gamma_1(X)\big)g^{-1}(\beta X) + \gamma_0(x) \tag{3}$$

Noting this equation for $P_T(Y = 1|X)$, he derived an equation for the naive estimator, β_1^*:

$$\beta_1^* = H(\beta_1) = g(P_T(Y = 1|X + 1)) - g(P_T(Y = 1|X)) \tag{4}$$

The actual equation for $H(\beta_1)$ is complex and non-intuitive, so he instead used Taylor approximations to obtain an estimate for this relationship between the true coefficients, β_1, and naive estimators, β_1^*. This is consistent in simple cases of misclassifcation, where misclassification is dependent only on binary variables with known effects. As well as the rigidity in the required data structure, the Neuhaus method also relies heavily on the assumptions of the distribution. While Neuhaus's method laid the groundwork for modern misclassification adjustment methods, it is not practical for the cases of complex misclassification that are frequently encountered today, as in our example data. Since his paper in 1999, researchers have expanded upon his principles to provide more flexible and nonparametric methods for misclassification adjustment.

2.2 Misclassification SIMEX

Cook and Stefanski developed the SIMEX (Simulation-Extrapolation) method in their 1994 paper such that no distributional assumptions need to be made about the misclassified data. Overall, SIMEX exploits the relationship of the size of the measurement error in the outcome (σ_u^2) to the bias of the naive estimator(s). The method adds additional measurement error to the data, and measures the induced bias in relation to the variance of the error that was added. Once a trend between the error and biased estimates is established, it can be used to extrapolate back to the case of no error.

In their 2006 paper, Kuchenhoff, Mwalili, and Lesaffre expanded the SIMEX method to correct coefficient estimates in the presence of a misclassified binary response. They refer to the independent variable that is derived disregarding measurement error as the naive estimator. An extrapolation function, $\mathcal{G}(\sigma_u^2)$, is used to estimate the relationship between the naive estimators and the measurement error, $\beta^*(\sigma_u^2)$, and implies that in the presence of no error, the true estimator could be derived: $\mathcal{G}(0) = \beta$. The SIMEX method is rooted in the parametric approximation to this function: $\mathcal{G}(\sigma_u^2) \approx \mathcal{G}(\sigma_u^2, \Gamma)$, where Γ is the vector of parameter estimates for the assumed distribution of the extrapolation function. Common distributions used are linear, quadratic (the most common), and loglinear. For example, $\mathcal{G}(\sigma_u^2, \Gamma) = \gamma_0 + \gamma_1\sigma_u^2 + \gamma_2(\sigma_u^2)^2$ for a quadratic estimation.

Through simulation, the SIMEX method adds additional measurement error with variance $\lambda\sigma_u^2$ to the misclassified variable, making the total measurement error equal to $(1+\lambda)\sigma_u^2$. The method then reevaluates the regression to obtain a new set of naive coefficients with the further misclassified outcome. By repeating this simulation step for a fixed set of λ's, one is able to obtain a parametric approximation for $\hat{\Gamma}$, in $\mathcal{G}(\sigma_u^2, \Gamma)$. Then, the function $\mathcal{G}(\sigma_u^2, \hat{\Gamma}$ is extrapolated back to 0 to obtain the adjusted SIMEX estimator, defined by $\mathcal{G}(0, \hat{\Gamma})$. This is best illustrated by the graph in Figure 1:

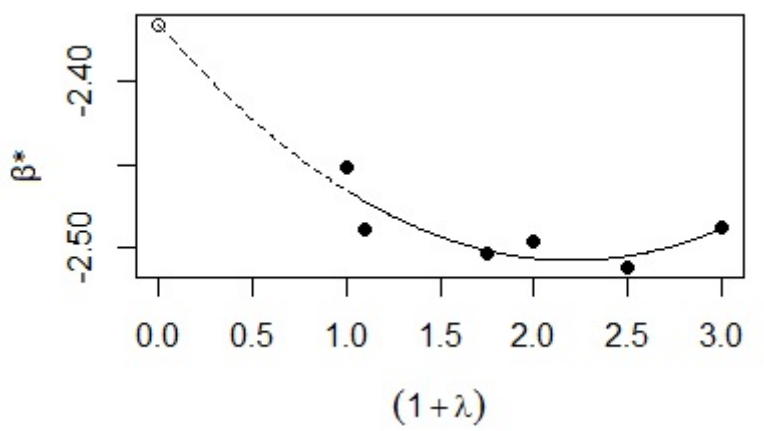

Figure 1: Proportion of Measurement Error $(1+\lambda)$ vs Naive Estimate (β^*) with the Superimposed Estimated Extrapolating Function, $\mathcal{G}(\sigma_u^2, \hat{\Gamma})$

Kuchenhoff, Mwalili, and Lesaffre discuss the calculation of the extrapolation function in terms of a linear model with a misclassifed binary covariate, the simplest case, and then apply those principles to the case of logistic regression with misclassified response. To expand the SIMEX method to the case of misclassification in a discrete response, Y, Kuchenhoff, Mwalili, and Lesaffre define the misclassification matrix Π as $\pi_{ij} = P(Y = i|T = j)$ (a $k \times k$ matrix, where k is the number of possible outcomes for T). In our example, the misclassification matrix is

$$\Pi = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix} = \begin{bmatrix} 1-\gamma_1(x) & \gamma_1(x) \\ 0 & 1 \end{bmatrix} \tag{5}$$

The misclassified estimator can then be defined as $\beta^*(\Pi^\lambda)$, and we can assume that when there is no misclassification, and thus $\Pi^0 = I_{k\times k}$, this equation gives the true value of β: $\beta^*(\Pi^0) = \beta$. In order for these functions to be well-defined, the determinant of Π must be greater than 0; in other words, both the specificity and sensitivity must be greater than 0.5. This is intuitive and corresponds to other research explaining that when the specificity or sensitivity are less than 50%, the data collection method performs worse than chance, rendering the data unusable. In practice, the equations for $\gamma_0(x)$ and $\gamma_1(x)$ are rarely known beforehand but may be estimated via a validation sample. However, it is prudent to note that the SIMEX method assumes that the sensitivity and specificity are known, so using estimated values could introduce further error.

The misclassification SIMEX (MC-SIMEX) algorithm begins by applying the misclassification matrix, Π^λ, to increase the bias in the response variable. Then, in the extrapolation step, the method uses the misclassified variables to extrapolate a parametric approximation:

$$\lambda \to \beta^*(\Pi^\lambda) \approx \mathcal{G}(1+\lambda, \Gamma) \tag{6}$$

The adjusted estimator $\hat{\beta}$ is then found by fitting the parametric approximation onto the set of extrapolated points, $[1+\lambda_k, \hat{\beta}(\lambda_k)]$, yielding an estimate of Γ. Using $\hat{\Gamma}$, you can obtain the MC-SIMEX estimator in the same way as the general SIMEX method:

$$\hat{\beta}_{MCSIMEX} := \mathcal{G}(0, \hat{\Gamma}) \tag{7}$$

Corresponding to $\lambda = -1$. The MC-SIMEX estimator will be consistent only when the extrapolation function is correct. The derivation of the extrapolation functions depends on the assumed distribution of $\mathcal{G}(\sigma_u^2, \hat{\Gamma})$ and can thus can be error-prone.

2.3 Maximum Likelihood

In their 2011 paper, Lyles, et al., introduced a more flexible method of misclassification adjustment that does not assume prior knowledge of the distribution for sensitivity and specificity. Their method proposed a data validation step in which a subset of the data is re-evaluated for the outcome with a more reliable data collection method. Using the more accurate second data collection method, they developed an approach using maximum likelihoods to adjust for misclassification. Developing a second logistic model for the association between the predictors and sensitivity/ specificity, they were able to account for differential misclassification using the validation sample:

$$\eta_t = logit[P(Y=1|T=t, X^*)] = \theta_0 + \theta_1 t + \sum \theta_i X_i^* \tag{8}$$

where X^* may be only a subset of the available covariates. Using this equation, sensitivity and specificity can be described as follows:

$$SE_{x_i} = P(Y=1|T=1, X=x_i) = \frac{exp(\eta_{1i})}{1+exp(\eta_{1i})} \tag{9}$$

$$SP_{x_i} = P(Y=0|T=0, X=x_i) = 1 - \frac{exp(\eta_{0i})}{1+exp(\eta_{0i})} \tag{10}$$

Since the specificity is fixed in our example, SP_{x_i} will be 1 for all values of x_i, and the parameters for SE_{x_i} can be estimated by a regression on Y with the validated outcome, T, and the covariates, X. The likelihood for the true parameters, L_T, is proportional to the product of the likelihood for the main data set, those not in the validation set, (L_m) and the likelihood using data only from those in the validation data set (L_v): $L_T \propto L_m \times L_v$. Due to the binary nature of the outcome, these likelihoods are well-defined.

$$L_m = \prod \big[(1 - SP_{x_i})P(T = 0|X = x_i) + SE_{x_i}P(T = 1|X = x_i)\big]^{y_i} \times \big[SP_{x_i}P(T = 0|X = x_i) + (1 - SE_{x_i})P(T = 1|X = x_i)\big]^{1-y_i} \tag{11}$$

$$L_v = \prod \big[SE_{x_j}P(T = 1|X = x_j)\big]^{y_j t_j} \times \big[(1 - SP_{x_j})P(T = 0|X = x_j)\big]^{y_j(1-t_j)} \times \big[(1 - SE_{x_j})P(T = 1|X = x_j)\big]^{(1-y_j)t_j} \times \big[SP_{x_j}P(T = 0|X = x_j)\big]^{(1-y_j)(1-t_j)} \tag{12}$$

An additional benefit to Lyles et al.'s method is the allowance for testing for completely non-differential misclassification through likelihood ratio tests. While the full likelihood, L_T does have to be explicitly programmed, once it is estabilished for the model, standard maximum likelihood estimation can be used to estimate the true values of β.

2.4 Multiple Imputation

In 2012, Edwards, et al. took a different approach in utilizing the validation sample to adjust for misclassification. They treated the misclassification in the outcome as a missing data problem where the only known values for the outcome were in the validation group; all other values of the outcome were treated as missing. Taking advantage of well-established methods for handling missing data, they performed multiple imputation on the outcome using the validation group to determine the relationship between the true outcome, T, the misclassified outcome, Y, and the covariates, X, forming an estimation to the equation:

$$logit\big(P(T = 1|Y, X)\big) = \alpha_0 + \alpha_1 Y + \boldsymbol{\alpha_2} Y X^* + \boldsymbol{\alpha_3} X) \tag{13}$$

Where X^* can be a subset of X or equivalent to X. In this way, the multiple imputation method lends itself nicely to handling differential as well as non-differential misclassification without prior knowledge of which type is expected, and even allows for the associations to be different between the specificity and the sensitivity through the interaction term $\boldsymbol{\alpha_2}$. Once this equation has been fitted, it can be used to multiply impute T in those not included in the validation sample.

For each of the K imputations, a regression is performed to obtain estimates for the coefficients. The final estimate for β is thus obtained from the average of all imputations:

$$\bar{\beta} = K^{-1} \sum_{k=1}^{K} \hat{\beta}^k \tag{14}$$

The variance of this new estimate is given by the sum of the average variance in the estimates and the mean squared error:

$$V(\bar{\beta}) = K^{-1} \sum_{k=1}^{K} V(\hat{\beta}^k) + (1 + K^{-1}) \left(\frac{1}{K - 1} \right) \sum_{k=1}^{K} (\hat{\beta}^k - \bar{\beta})^2 \tag{15}$$

This method has the added flexibility of being compatible with standard missing data methods and approaches asymptotic efficiency in the estimates as the value of K increases.

3 Simulation

In order to evaluate potential biases due to misclassification, I conduct a simulation that resembles the mechanism of misclassification in the MCHDCS study. To simplify the model, I employ only two independent covariates: one continuous, $x_c \sim N(0,1)$, and one binary, $x_b \sim Bern(0.4)$. The true outcome, T, and the probability of misclassification are then generated from a model of these values.

To explore different intensities and directions of differential misclassification, I vary the coefficients for x_c and x_b but fix the probability of misclassification by adjusting the intercept value accordingly. I then generate two additional outcome variables that are misclassified differentially and non-differentially. I range the probability of misclassification from 0 to 50%, since data with a suspected misclassification rate above 50% is considered heavily faulty data and should not be analysed, but rather re-collected (Neuhaus 1999). This is because a suspected misclassification rate above this threshold is considered worse than misclassification at random.

3.1 Set-up

For each simulation, I generate a data set of n = 1,000,000 observations with x_c (continuous) and x_b (binary) under the previously specified models and use them to create the data generating model:

$$logit(P(T=1)) = -1.5 + 0.5x_c + 0.5x_b \tag{16}$$

I then generate two misclassified outcomes, Y_d and Y_{nd}, that are differentially and non-differentially misclassified, respectively. In order to emulate the model from the example data, I misclassify the outcome only when $T = 1$ and make no changes when $T = 0$, maintaining the specificity at 100% and thus $\gamma_0 = 0$. The mean sensitivity is varied from 50% to 100% in both outcomes. The non-differentially misclassified variable, Y_{nd}, is generated by simply switching the value of T from 1 to 0 with the probability $\gamma_1 \in [0, 0.5]$. The differentially misclassified variable, Y_d, is switched from 1 to 0 with a probability, $\gamma_1(x)$, that is modeled by the equation:

$$logit(\gamma_1(x)) = d_0 + d_c x_c + d_b x_b \tag{17}$$

I vary the d_c and d_b coefficients to explore the effect of the association between the marginal probability of misclassification and the explanatory variables. Note that $(d_c, d_b) = (0, 0)$ implies non-differential misclassification. To facilitate the comparison of the effects of non-differential versus differential misclassification, I adjust the d_0 coefficient to maintain $E(\gamma_1(x)) = \gamma_1$.

3.2 Evaluating Induced Biases

I will first examine the biases when misclassification is dependent only on one of the variables in the model to evaluate how this would effect the entire model. I do it for both the continuous and binary variables separately to explore whether they have different influences on the bias produced. The marginal probability of misclassification in the differential case is thus modeled by the two equations:

Scenario 1	Scenario 2
$logit(\gamma_1(x)) = d_{00} + x_b$	$logit(\gamma_1(x)) = d_{01} + x_c$

Note that in each equation the coefficient for the covariate of interest in the misclassification model is set to 1 and the other coefficient is set to 0. The intercept coefficient will vary and is not necessarily representing the same value in both equations.

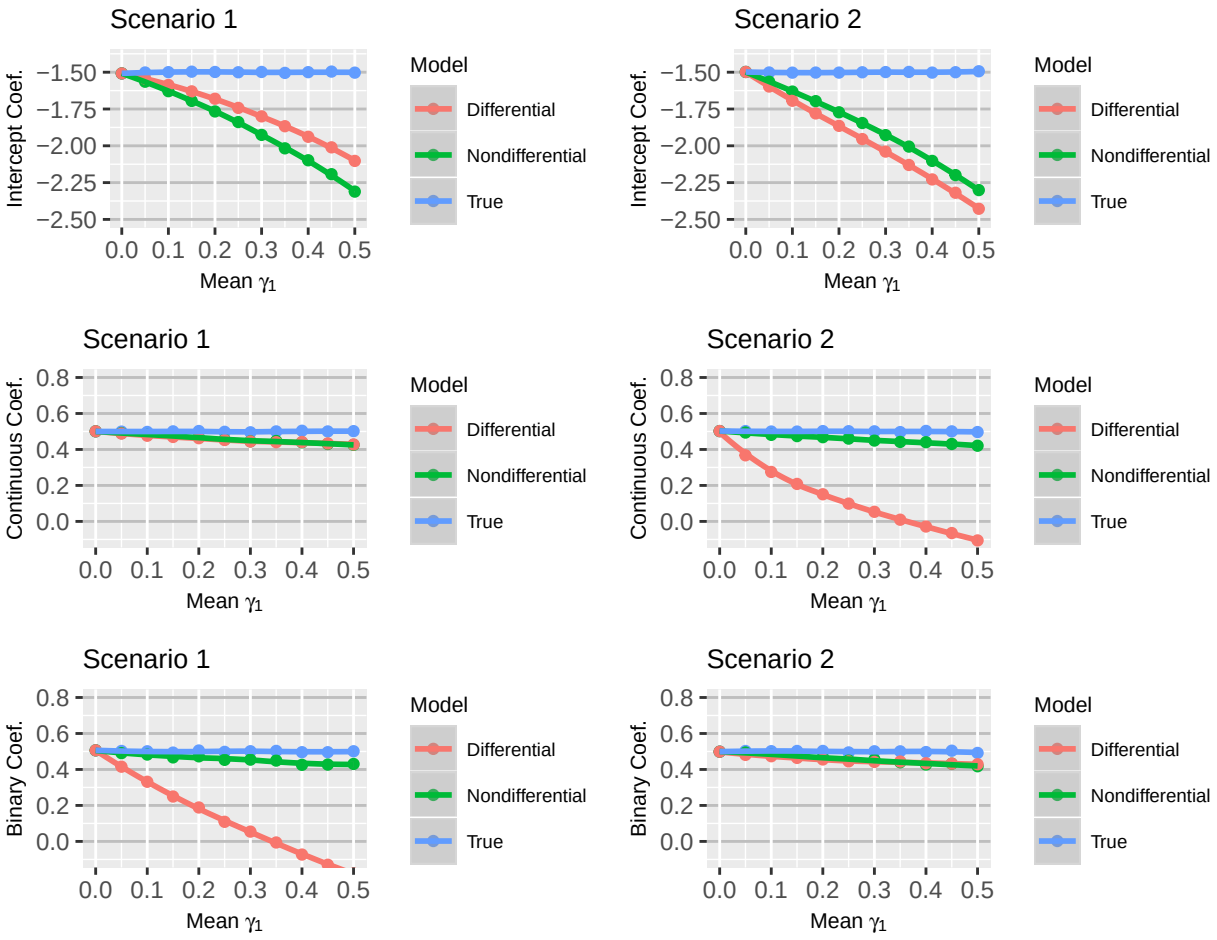

Figure 2: Univariate Misclassification, Binary vs. Continuous

Above: Left: differential model dependent only on the binary variable. Right: differential model dependent only on the continuous variable. The derived coefficients for the model are plotted with the corresponding marginal probability of misclassification.

The plots in Figure 2, from both Scenario 1 and Scenario 2, illustrate that when misclassification is not dependent on a particular variable, the bias induced on that variable is equivalent to the bias from a totally non-differential model. This is true for both continuous and binary variables. The bias induced in the differential model remains isolated to the variable(s) affecting misclassification.

One troubling feature to note is what happens as the probability of misclassification increases. Due to the coefficient-dependent misclassification, the coefficient determined by the regression trends towards 0 and can even switch signs if the misclassification rate is large, seen in probabilities of misclassification >0.35 for the binary coefficient in Scenario 1 and for the continuous coefficient in Scenario 2.

The bias from non-differential misclassification is small in the coefficient estimates and appears to only heavily affect the intercept term. Since the bias trends are similar in the intercept term no matter what kind of misclassification is present, I will not present the graphs for the intercept in the following examples, as they are remarkably similar to the intercept graphs above and provide little further inference.

In the following simulations, I evaluated the effect of having a negative correlation between the misclassification and our variable of interest to confirm that the magnitude of the effect is in fact similar to the positive correlation model. I also evaluate misclassification due to multiple variables to see if there is any compounding

effect. The probability of misclassification in the differential case is thus modeled by the two equations:

Scenario 3	Scenario 4
$logit(\gamma_1(x)) = d_{00} - x_c$	$logit(\gamma_1(x)) = d_{01} + x_c + x_b$

Note that the intercept coefficient will vary and is not necessarily representing the same value in both equations.

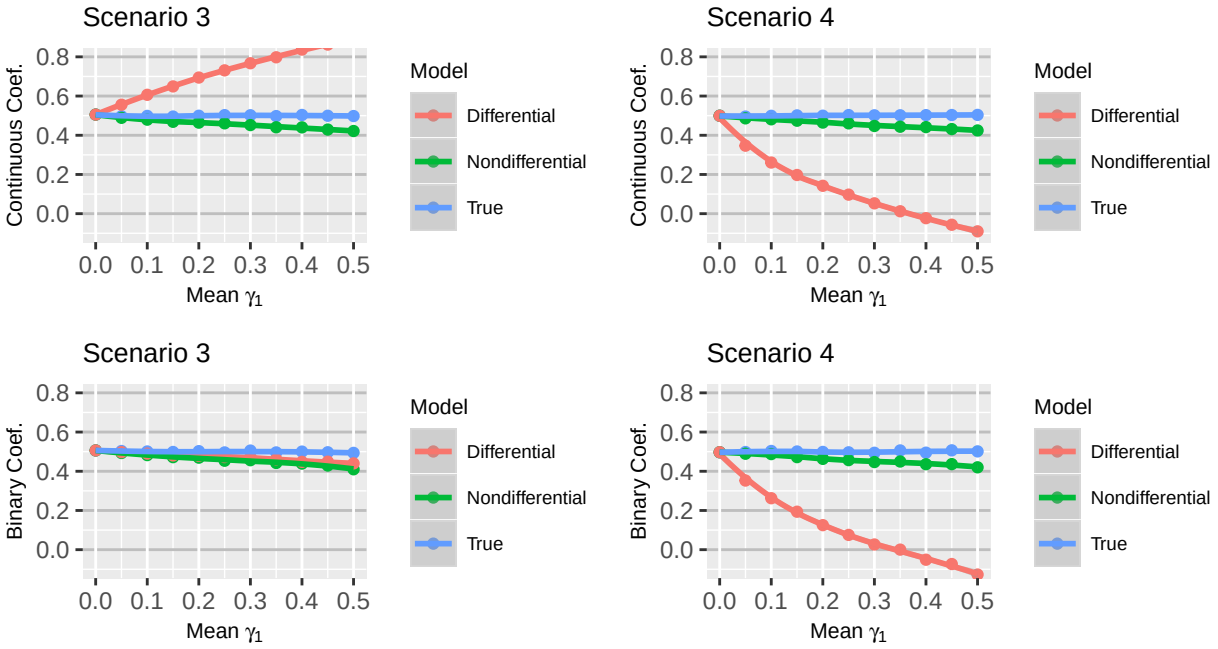

Figure 3: Negative and Multivariate Misclassification Models

Above: Left: differential model with a negative coefficient for x_c in the misclassification model. Right: differential model dependent on both continuous and binary variables. The derived coefficients for the model are plotted with the corresponding probability of misclassification.

The simulation with a negative coefficient for differential misclassification behaves as expected, mirrored across the true value of the coefficient. This could be problematic as the bias trends the coefficient estimates away from zero, possibly causing variables to appear more strongly associated with the outcome than they truly are and thus leading to incorrect inferences. We can also see that the effects of misclassification on a coefficient do not compound when the misclassification is due to multiple variables. The coefficients are similarly biased in the model with both variables as they are in the model where misclassification was only due to one variable.

In the next set of simulations, I evaluated the effect of having an association between the misclassification and the variables of interest at more extreme ends of the spectrum, where the association is particularly small or large. Since the effects of differential misclassification are similar between continuous and binary variables, I will present only the continuous variable. The probability of misclassification in the differential case is thus modeled by the two equations:

Scenario 5	Scenario 6
$logit(\gamma_1(x)) = d_{00} + 0.25x_c$	$logit(\gamma_1(x)) = d_{01} + 1.5x_c$

Note that the intercept coefficient will vary and is not necessarily representing the same value in both equations.

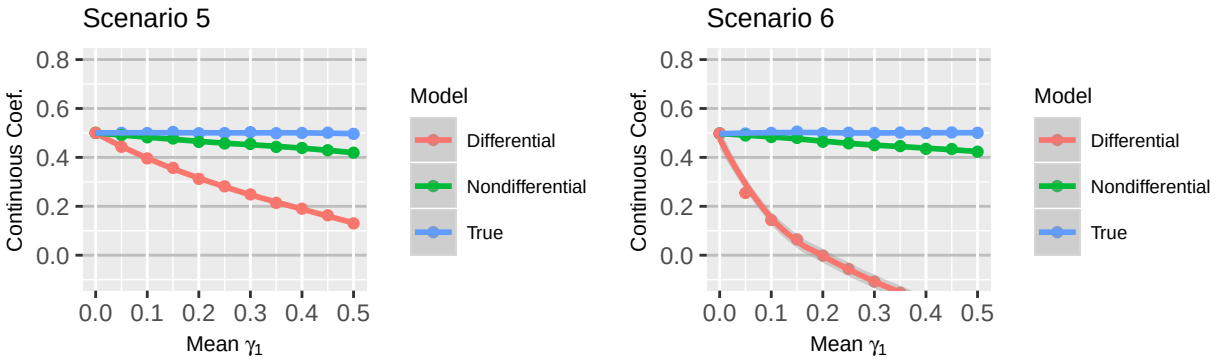

Figure 4: Varying Coefficients of Misclassification Models

Above: Left: differential model with $d_c = 0.25$. Right: differential model with $d_c = 1.5$.

The graphs in Figure 4 are relatively intuitive. With a smaller magnitude, the regression is less affected by the misclassification, but could still drive coefficient estimates away from the true value. The case with the larger magnitude is much more concerning. With a probability of misclassification at just 0.1, the coefficient estimate has already switched signs. This means that even moderately misclassfied data could bias results such that they present the direct opposite of the truth. This has strong implications when it is believed that the misclassification is highly dependent on the covariate of interest.

3.3 Evaluating Adjustment Methods

In order to demonstrate the capabilities of the presented adjustment methods, I generated data using similar coefficients to our example study data with the equation:

$$logit\big(P(T = 1|X)\big) = -1.5 + 0.5x_1 + 0.15x_2 - 0.4x_3 - 0.4x_4 \tag{18}$$

Where (x_1, x_2, x_3) are N(0,1) and x_4 is Bin(0.17) to resemble the example data. I performed 5,000 replications with n = 3000 subjects simulated for each one. After generating, I misclassified the data, holding specificity = 1 and varying the sensitivity dependent on a subset of the continuous variables to replicate our unique case of non-differential misclassification.

$$\gamma_0(X) = 1 \qquad \gamma_1(X) = \frac{exp(\theta_0 - 0.65x_2 + 0.4x_3)}{1 + exp(\theta_0 - 0.65x_2 + 0.4x_3))} \tag{19}$$

This emphasizes that misclassification and the outcome are not always dependent on the same set of variables; misclassification can be due to all of the variables of interest, a subset of them, or none of them. To illustrate the effect that varying degrees of sensitivity can have on these methods, I ran one set of the simulations for $\theta_0 = -1$ and the other half for $\theta_0 = -0.2$, which corresponds to an average sensitivity of approximately 0.73 and 0.55 respectively. The former is intended to be consistent with our example case, and the latter explores a marginally acceptable sensitivity. In addition to varying the sensitivity, I also run each method (and sensitivity) for three validation group sizes to explore the necessary sample size for these validation-based methods to be effective: 500, 1000, and 1500.

Table 1: Assessment of Adjustment Methods on X_1: True Value = 0.5

	Val. Size = 500				Val. Size = 1000				Val. Size = 1500			
Method	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE
Avg. Sensitivity of 0.55												
Naive	0.4502	0.0619	0.0615	0.0063	0.4502	0.0619	0.0615	0.0063	0.4502	0.0619	0.0615	0.0063
MLE	0.5082	0.0952	0.0662	0.0091	0.5029	0.0701	0.0614	0.0049	0.5026	0.0612	0.0577	0.0038
Imputed	0.4815	0.0950	0.0938	0.0094	0.4885	0.0710	0.0707	0.0052	0.4919	0.0627	0.0612	0.0040
SIMEX	0.4585	0.0609	0.0660	0.0054	0.4656	0.0595	0.0705	0.0047	0.4742	0.0580	0.0742	0.0040
Avg. Sensitivity of 0.73												
Naive	0.4699	0.0558	0.0559	0.0040	0.4699	0.0558	0.0559	0.0040	0.4699	0.0558	0.0559	0.0040
MLE	0.5029	0.0790	0.0590	0.0062	0.5029	0.0628	0.0566	0.0040	0.5023	0.0568	0.0547	0.0032
Imputed	0.4798	0.0805	0.0797	0.0069	0.4839	0.0634	0.0630	0.0043	0.4884	0.0578	0.0569	0.0035
SIMEX	0.4765	0.0554	0.0584	0.0036	0.4829	0.0555	0.0607	0.0034	0.4895	0.0548	0.0626	0.0031

Table 2: Assessment of Adjustment Methods on X_2: True Value = 0.15

	Val. Size = 500				Val. Size = 1000				Val. Size = 1500			
Method	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE
Avg. Sensitivity of 0.55												
Naive	0.4254	0.0605	0.0613	0.0795	0.4254	0.0605	0.0613	0.0795	0.4254	0.0605	0.0613	0.0795
MLE	0.1469	0.1179	0.0647	0.0139	0.1493	0.0711	0.0595	0.0051	0.1508	0.0595	0.0558	0.0035
Imputed	0.1582	0.0871	0.0898	0.0076	0.1563	0.0683	0.0681	0.0047	0.1538	0.0597	0.0592	0.0036
SIMEX	0.3833	0.0602	0.0649	0.0580	0.3336	0.0593	0.0700	0.0372	0.2737	0.0577	0.0743	0.0186
Avg. Sensitivity of 0.73												
Naive	0.3330	0.0558	0.0551	0.0366	0.3330	0.0558	0.0551	0.0366	0.3330	0.0558	0.0551	0.0366
MLE	0.1502	0.0991	0.0581	0.0098	0.1506	0.0654	0.0553	0.0043	0.1517	0.0565	0.0531	0.0032
Imputed	0.1627	0.0745	0.0799	0.0057	0.1591	0.0589	0.0623	0.0036	0.1575	0.0534	0.0557	0.0029
SIMEX	0.2983	0.0548	0.0568	0.0250	0.2581	0.0539	0.0597	0.0146	0.2136	0.0536	0.0621	0.0069

Table 3: Assessment of Adjustment Methods on X_3: True Value = -0.4

	Val. Size = 500				Val. Size = 1000				Val. Size = 1500			
Method	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE
Avg. Sensitivity of 0.55												
Naive	-0.5350	0.0625	0.0621	0.0221	-0.5350	0.0625	0.0621	0.0221	-0.5350	0.0625	0.0621	0.0221
MLE	-0.4214	0.1289	0.0656	0.0171	-0.4035	0.0786	0.0605	0.0062	-0.4022	0.0628	0.0569	0.0039
Imputed	-0.3937	0.0874	0.0890	0.0077	-0.3975	0.0679	0.0679	0.0046	-0.3977	0.0589	0.0597	0.0035
SIMEX	-0.5144	0.0607	0.0666	0.0168	-0.4885	0.0601	0.0713	0.0114	-0.4578	0.0577	0.0752	0.0067
Avg. Sensitivity of 0.73												
Naive	-0.4930	0.0560	0.0561	0.0118	-0.4930	0.0560	0.0561	0.0118	-0.4930	0.0560	0.0561	0.0118
MLE	-0.4208	0.1010	0.0586	0.0106	-0.4085	0.0692	0.0560	0.0049	-0.4051	0.0580	0.0540	0.0034
Imputed	-0.3988	0.0753	0.0755	0.0057	-0.3986	0.0605	0.0612	0.0037	-0.3983	0.0552	0.0559	0.0030
SIMEX	-0.4763	0.0552	0.0583	0.0089	-0.4552	0.0548	0.0607	0.0060	-0.4335	0.0543	0.0627	0.0041

Table 4: Assessment of Adjustment Methods on X_4: True Value = -0.4

	Val. Size = 500				Val. Size = 1000				Val. Size = 1500			
Method	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE	Avg(Est)	SD(Est)	Avg(SE(Est))	MSE
Avg. Sensitivity of 0.55												
Naive	-0.3738	0.1791	0.1758	0.0328	-0.3738	0.1791	0.1758	0.0328	-0.3738	0.1791	0.1758	0.0328
MLE	-0.4022	0.2442	0.1809	0.0596	-0.4043	0.1900	0.1688	0.0361	-0.4070	0.1683	0.1596	0.0284
Imputed	-0.3797	0.2248	0.2348	0.0509	-0.3844	0.1787	0.1842	0.0322	-0.3899	0.1612	0.1645	0.0261
SIMEX	-0.3857	0.1768	0.1881	0.0314	-0.3884	0.1703	0.1997	0.0291	-0.3926	0.1671	0.2087	0.0280
Avg. Sensitivity of 0.73												
Naive	-0.3821	0.1566	0.1580	0.0248	-0.3821	0.1566	0.1580	0.0248	-0.3821	0.1566	0.1580	0.0248
MLE	-0.3889	0.1994	0.1621	0.0399	-0.3990	0.1672	0.1560	0.0279	-0.4052	0.1536	0.1512	0.0236
Imputed	-0.3693	0.1978	0.2064	0.0400	-0.3801	0.1621	0.1679	0.0267	-0.3903	0.1524	0.1550	0.0233
SIMEX	-0.3928	0.1543	0.1638	0.0238	-0.3951	0.1520	0.1693	0.0231	-0.3998	0.1515	0.1734	0.0229

Tables 1 and 4 illustrate that when the misclassification is not dependent on the variable, the coefficient estimates for that variable from the naive and adjusted models are all very similar and close to the true value. Subsequently the MSE does not decrease significantly in the adjusted models over the naive model. However, the adjustment methods have a large impact on the coefficients that were involved in the misclassification, x_2 and x_3, which is congruent with what was observed in the simulations. With both variables, the misclassification induces a bias away from the null. While misclassification was dependent on both x_2 and x_3, the association was stronger with x_2; subsequently, we see more error in this coefficient. The maximum likelihood and multiple imputation adjustment methods are effective at bringing the estimates closer to the true values, while the MC-SIMEX method underperforms at all validation levels.

From Tables 1-4, we can see that the maximum likelihood and multiple imputation methods outperform the MC-SIMEX method in almost every case. The MC-SIMEX method was developed for cases of known, rather than estimated, sensitivity and specificity, as is the case with the estimates gathered from the validation groups in the simulations. Subsequently, the MC-SIMEX method is able to produce better adjustments with large validation sizes when the mechanism for misclassification can be better estimated and more true data is available; however, such large validation sizes are rarely possible in practice, and the other methods produce estimates closer to the true values regardless of validation size. Even with the smallest validation size, the maximum likelihood and multiple imputation methods are able to estimate the true values with much less error. The multiple imputation method appears to perform the best out of the three, but the margin of improvement between this method and the maximum likelihood is small. However, by comparing the SD(Est) and the Avg(SE(Est)), one can observe that the maximum likelihood method also underestimates the uncertainty in the data, which can lead to false-positives in the final analyses.

4 Data Analysis

4.1 Data Introduction

The data collected from the Mid-South Coronary Heart Disease Cohort Study (MCHDCS) resulted in 488 cases of hospitalization during the follow-up period after the data was expanded to include the Vanderbilt Health Affiliated Network (VHAN) as well as Vanderbilt Medical Center. We evaluate covariates hypothesized to affect the hospitalization rates that include: age, sex, objective health, subjective health, race, education, trust in healthcare, and access to healthcare. Objective health is quantified through the patient's Elixhauser scores, calculated from their ICD codes in the data base. In order to adjust for some patients having more extensive ICD records, the number of years spanning ICD codes for each patient was included in the model. Trust in healthcare is a four-level variable based on survey data regarding a patient's trust in doctors, other healthcare providers (OHP), both doctors and OHP, or neither. Trust in both is used as the reference. Access to healthcare is a continuous variable, ranging from 0 to 3, derived from averaging 2 measures of access to healthcare based on finances and doctor/office availability. A high value in this variable indicates better access to healthcare. The variable of interest, health literacy, was treated as a continuous variable ranging from 3 to 15, with a high value reflecting high competency. Table 5 describes the covariates in the data set divided by hospitalization status and location.

From Table 5, significant differences are observed in the marginal comparisons of the true hospitalized patients versus those not hospitalized in all covariates except ICD observation time, trust in healthcare, race, and gender. Those hospitalized tend to be older and in worse health, objectively and subjectively. This is intuitive as it would be expected for those groups to be hospitalized at a higher rate. Interestingly, those hospitalized tend to be lower health literacy and fewer years of education. This could indicate that those who understand their health less are more likely to go to the hospital when necessary. However, this only gives preliminary information as these associations are likely to change in the full, adjusted model.

Table 5: Comparison of Demographics Across Outcome with N_n=2117, N_v=364, and N_e=124, respectively

		Hospitalizations		
Variable	Missing	None	VUMC Only	Elsewhere
Age	1	68 (10.5)	68.5 (10.6)	72 (12)
Weight Total Accessibility	51	1.3 (0.3)	1.4 (0.4)	1.3 (0.3)
Mean of Subjective Health Measures	0	3.2 (0.8)	2.8 (0.8)	2.8 (0.8)
ICD Collection Time (years)	0	9.4 (5)	10.4 (5.4)	8.5 (5)
Elixhauser Score	0	11.6 (10.8)	20.8 (12.6)	13.6 (10.4)
Scaled Sum of Health Literacty Scores	7	12.6 (2.9)	12.3 (3.1)	11.6 (3.4)
Trust in Healthcare	47			
Trust in Both		1463 (69%)	252 (69%)	86 (69%)
Trust Doctors Only		376 (18%)	54 (15%)	16 (13%)
Trust OHP Only		28 (1%)	3 (1%)	4 (3%)
Don't Trust Either		215 (10%)	45 (12%)	16 (13%)
Race, binary	0			
White		1890 (89%)	307 (84%)	113 (91%)
Other		227 (11%)	57 (16%)	11 (9%)
Gender	0			
Male		1446 (68%)	251 (69%)	83 (67%)
Female		671 (32%)	113 (31%)	41 (33%)
Education	37			
Did not graduate high school		190 (9%)	43 (12%)	15 (12%)
High school graduate or GED		516 (24%)	88 (24%)	34 (27%)
Some college or 2-year degree		579 (27%)	117 (32%)	34 (27%)
College graduate		365 (17%)	43 (12%)	15 (12%)
More than a college degree		441 (21%)	67 (18%)	21 (17%)

4.2 Investigation into Misclassification

After data collection was expanded to include the surrounding hospitals, a 34.1% increase in the number of patients that were hospitalized was observed. The table below illustrates the misclassification of the outcome.

Table 6: Comparison of Newly Added Hospitalization Data

		VUMC Only		
		Not Hospitalized	Hospitalized	Total
All Hospitals	Not Hospitalized	2117	0	2117
	Hospitalized	124	364	488
	Total	2241	364	2605

From this table, we can see that we have a sensitivity of 74.6% and a specificity of 100%. The sensitivity and specificity can be used to adjust for the misclassification and generate unbiased estimates. However, first, the type of misclassification must be investigated. I perform a logistic regression on the 488 cases with the outcome being an binary indicator of misclassification in our original data collection. I use the same covariates from Table 5. The results from this regression are in Table 7 below.

Table 7: Regression Results for Predicting Misclassification

Variable	DF	Odds Ratios	P-value
Health Literacy	1		0.0109
		0.75 (0.6, 0.93)*	
Elixhauser Score	1		<0.0001
		0.52 (0.41, 0.67)*	
Collection Time	1		0.00486
		0.7 (0.55, 0.9)*	
Subjective Health	1		0.429
		0.9 (0.7, 1.17)	
Trust in Healthcare	3		0.633
Trust Doctors Only		1.13 (0.58, 2.19)	
Trust OHP Only		2.77 (0.58, 13.18)	
Don't Trust Either		1.05 (0.52, 2.09)	
Weighted Access	1		0.131
		0.83 (0.66, 1.06)	
Race, binary	1		0.263
Other		0.65 (0.31, 1.38)	
Gender	1		0.318
Female		1.28 (0.79, 2.09)	
Age	1		0.00363
		1.44 (1.13, 1.84)*	
TOTAL	11		<0.0001

[a] All continuous variables are standardized.

The regression indicates that the misclassification observed in the outcome is dependent on health literacy, Elixhauser score, ICD observation time, and age; i.e., the misclassification is differential and thus non-random. Patients who were less health literate, in worse health, observed for a shorter amount of time, and/or older were more likely to be hospitalized somewhere other than Vanderbilt Medical Center and thus more likely to be misclassified by the original data collection. This regression will be used to estimate the model for the misclassification in the data, demonstrating the ability of the adjustment methods to alleviate this problem.

4.3 Data Analysis

As displayed in the simulation studies in Section 3.2, differential misclassification can have a strong impact on coefficient estimates. To display how the misclassification can bias the results of the regression in typical analyses, I present the true and naive regressions in Table 8: one with the original misclassified outcome and one with the true outcome after further data collection.

Table 8: Regression Results for Misclassified vs. True Model

		Misclassified		True	
Variable	DF	Odds Ratios	P-value	Odds Ratios	P-value
Health Literacy	1		0.00286		0.0397
		1.22 (1.07, 1.4)*		1.13 (1.01, 1.26)*	
Elixhauser Score	1		<0.0001		<0.0001
		1.91 (1.7, 2.15)*		1.66 (1.5, 1.84)*	
Collection Time	1		0.11		0.919
		1.1 (0.98, 1.24)		1.01 (0.9, 1.12)	
Subjective Health	1		<0.0001		<0.0001
		0.68 (0.59, 0.79)*		0.66 (0.58, 0.74)*	
Trust in Healthcare	3		0.182		0.0826
Trust Doctors Only		0.7 (0.5, 0.99)*		0.69 (0.51, 0.93)*	
Trust OHP Only		0.67 (0.2, 2.3)		1.2 (0.5, 2.86)	
Don't Trust Either		1.07 (0.73, 1.55)		1.06 (0.76, 1.47)	
Weighted Access	1		0.00274		0.0241
		1.2 (1.07, 1.36)*		1.13 (1.02, 1.26)*	
Race, binary	1		0.106		0.225
Other		1.34 (0.94, 1.92)		1.22 (0.88, 1.69)	
Gender	1		0.0266		0.0689
Female		0.74 (0.56, 0.97)*		0.8 (0.64, 1.02)	
Age	1		0.453		0.00387
		1.05 (0.92, 1.19)		1.18 (1.05, 1.32)*	
TOTAL	11		<0.0001		<0.0001

[a] All continuous variables are standardized.

After the second round of data collection where more cases were found, gender no longer had a significant effect, and age became significant. The magnitude of the effects for health literacy, Elixhauser score, access, and gender were mitigated and brought closer to the null, while other variables were more significant in the true model. This could mislead researchers and hinder reproducibility of the results.

Having demonstrated the gravity of misclassified data on the results, I then employed the misclassification adjustment methods in attempt to adjust for the measurement error in the data. I randomly sampled 1000 patients from the data set to mimic a validation sample for the adjustment methods. In practice, only the patients that had not been hospitalized in the original data set would need to be contacted, since the specificity is 1, which would be around 860 patients on average in the example data, based on the rate of hospitalization in the original data. I repeated the validation sampling and reapplied the adjustments methods 100 times and averaged the results in the table below. The methods are presented with the coefficients and subsequent 95% confidence intervals for each variable and method combination. An asterisk next to the confidence interval indicates that the value is significantly different than the null based on the variance from that model. Since the true and naive models do not involve the validation samples, the calculation was only performed once and those values are presented along side the averages from the adjustment methods for comparison.

Table 9: Regression Results for Model Comparison

	Truth	Naive	Mult. Impute	MLE	MCSimex
Health Literacy	1.13 (1.01, 1.26)*	1.22 (1.07, 1.4)*	1.1 (0.95, 1.27)	1.14 (1, 1.3)*	1.22 (1.06, 1.41)*
Elixhauser Score	1.66 (1.5, 1.84)*	1.91 (1.7, 2.15)*	1.68 (1.5, 1.89)*	1.67 (1.49, 1.87)*	1.92 (1.7, 2.16)*
ICD Collection Time	1.01 (0.9, 1.12)	1.1 (0.98, 1.24)	1.02 (0.9, 1.15)	1.02 (0.91, 1.14)	1.1 (0.98, 1.24)
Subjective Health	0.66 (0.58, 0.74)*	0.68 (0.59, 0.79)*	0.72 (0.63, 0.82)*	0.66 (0.57, 0.75)*	0.68 (0.59, 0.79)*
Trust: Trust Doctors Only	0.69 (0.51, 0.93)*	0.7 (0.5, 0.99)*	0.75 (0.54, 1.04)	0.7 (0.5, 0.98)*	0.7 (0.5, 0.99)*
Trust: Trust OHP Only	1.2 (0.5, 2.86)	0.67 (0.2, 2.3)	0.91 (0.31, 2.65)	0.93 (0.31, 2.79)	0.66 (0.17, 2.55)
Trust: Don't Trust Either	1.06 (0.76, 1.47)	1.07 (0.73, 1.55)	1.05 (0.73, 1.51)	1.08 (0.74, 1.56)	1.06 (0.73, 1.53)
Weighted Access	1.13 (1.02, 1.26)*	1.2 (1.07, 1.36)*	1.15 (1.02, 1.29)*	1.16 (1.03, 1.31)*	1.2 (1.06, 1.35)*
Race (binary): Other	1.22 (0.88, 1.69)	1.34 (0.94, 1.92)	1.25 (0.88, 1.78)	1.28 (0.9, 1.83)	1.35 (0.94, 1.94)
Sex: Female	0.8 (0.64, 1.02)	0.74 (0.56, 0.97)*	0.81 (0.63, 1.05)	0.78 (0.6, 1.01)	0.74 (0.56, 0.97)*
Age Divided by 5	1.18 (1.05, 1.32)*	1.05 (0.92, 1.19)	1.16 (1.01, 1.33)*	1.18 (1.04, 1.34)*	1.05 (0.93, 1.19)

Interestingly, in our complex example of misclassification with a relatively small validation size, the MC-SIMEX method was unable to produce estimates better than the naive model. However, the maximum likelihood and multiple imputation methods strongly reduced the error in the coefficients. In particular, the MCHDCS study was interested in the health literacy effect which the maximum likelihood method captured well. The multiple imputation method produced a closer estimate to the true value; however, unlike the true model, it did not indicate that the effect was significantly different than the null. The multiple imputation method proves to be more conservative than the maximum likelihood method in the trust coefficient as well. However, having learned from the simulations that the maximum likelihood tends to underestimate uncertainty, its tight confidence intervals may not actually be representative of the data.

5 Discussion

Non-differential, or random, misclassification is present in small quantitites in most data collection, and the simulation studies reviewed here illustrated that this type only mildly biases the coefficient estimates toward the null. While this could potentially cause significant associations to be missed in cases of heavy misclassification, the true problem arises when it is suspected that the probability of misclassification is dependent on the variables in the model, as was discovered in the MSCHDCS data. This differential misclassification can have a large impact on the validity of the results. In the MSCHDCS data, gender was erroneously significant in the naive model, and the significance of age was missed altogether. Since the misclassification was dependent on the target variable, health literacy, adjustment was even more prudent to our results, and the magnitude of the health literacy coefficient was reduced in the adjusted models. In contrast, there was an increase in the age coefficient as it became significant after adjustment. This illustrates how the direction of the bias in differential misclassification is not consistent and thus not easily predicted.

The maximum likelihood method proves to be effective in handling the case of differential misclassification in both the example data and the simulations. However, the performance of both the multiple imputation and maximum likelihood adjustment methods is similar. This is intuitive since they both draw on likelihood properties. While the MC-SIMEX method is appealing due to the complex derivation and multiple simulations, it was not able to provide results better than the naive model when utilizing an estimated model for the complex mechanism of misclassificaiton. The MC-SIMEX method would be best reserved for cases where the sensitivity and specificity are known and more straightforward.

While it will rarely be possible to identify the model for differential misclassification without having prior knowledge, the model can be estimated with a validation group and still provide consistent results. The simulation studies highlighted that estimates can be greatly improved with these methods even when the validation size is small. This has great impact on research, since collecting validation samples can be labor- and cost-intensive, but studies can derive relatively unbiased results using the maximum likelihood or multiple imputation methods with modest validation sizes. The multiple imputation method may be favorable since the technique for applying this method is well-established in most statistical coding languages, but it may underreport significant associations if they are small in magnitude.

References

Cook, J. R., & Stefanski, L. A. (1994) Simulation-Extrapolation Estimation in Parametric Measurement Error Models. Journal of the American Statistical Association, 89(428), 1314-28.

Edwards, J. K., Cole, S. R, Troester, M. A., & Richardson, D. B. (2013). Accounting for Misclassified Outcomes in Binary Regression Models Using Multiple Imputation With Internal Validation Data. American Journal of Epidemiology, 177(9), 904-12.

Kuchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. Biometrics, 62(1), 85-96.

Luan, X., Pan, W., Gerberich, S. G., & Carlin, B. P. (2005). Does it always help to adjust for misclassification of a binary outcome in logistic regression? Statistics in Medicine, 24(14), 2221-34.

Lyles, R. H., Tang, L., Superak, H. M., King, C. C., Celentano, D. D., Lo, Y., & Sobel, J. D. (2011). Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. Epidemiology (Cambridge, Mass.), 22(4), 589-97.

Magder, L. S., & Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. American Journal of Epidemiology, 146(2), 195-203.

Mwalili, S. M. (2009). Bayesian and Frequentist Approaches to Correct for Misclassification Error with Applications to Caries Research. Katholieke Universiteit Leuven.

Neuhaus, J. M. (1999). Bias and Efficiency Loss Due to Misclassified Responses in Binary Regression. Biometrika, 86(4), 843-55.