Statistics in Medicine

Research Article

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Corrected ROC Analysis for Misclassified Binary Outcomes

Matthew Zawistowski[a,b]*, Jeremy B. Sussman[a,c], Timothy P. Hofer[a,c], Douglas Bentley[a], Rodney A. Hayward[a,c], Wyndy L. Wiitala[a]

1. Abstract

Creating accurate risk prediction models from Big Data resources such as Electronic Health Records (EHRs) is a critical step toward achieving precision medicine. A major challenge in developing these tools is accounting for imperfect aspects of EHR data, particularly the potential for misclassified outcomes. Misclassification, the swapping of case and control outcome labels, is well-known to bias effect size estimates for regression prediction models. In this paper, we study the effect of misclassification on accuracy assessment for risk prediction models and find that it leads to bias in the Area Under the Curve (AUC) metric from standard ROC analysis. The extent of the bias is determined by the false positive and false negative misclassification rates as well as disease prevalence. Notably, we show that simply correcting for misclassification while building the prediction model is not sufficient to remove the bias in AUC. We therefore introduce an intuitive misclassification-adjusted ROC procedure that accounts for uncertainty in observed outcomes and produces bias-corrected estimates of the true AUC. The method requires that misclassification rates are either known or can be estimated, quantities typically required for the modeling step. The computational simplicity of our method is a key advantage, making it ideal for efficiently comparing multiple prediction models on very large datasets. Finally, we apply the correction method to a hospitalization prediction model from a cohort of over one million patients from the Veterans Health Administrations EHR. Implementations of the ROC correction are provided for Stata and R.

2. Introduction

Predicting a binary outcome using a set of covariates is common practice in many areas of research[1, 2]. Typically this involves fitting a predictive model on observed data, then assessing how accurately prediction probabilities from that model discriminate cases from controls. A standard method for quantifying prediction accuracy is to compute AUC, the Area Under the Receiver Operating Characteristic (ROC) Curve[3]. The ROC curve plots the sensitivity (true positive) and specificity (true negative) of the prediction model at potential discrimination thresholds and the AUC is the probability

[a] *Veterans Affairs Center for Clinical Management Research, Ann Arbor, MI 48105, USA*
[b] *Department of Biostatistics, University of Michigan, Ann Arbor, MI, 48109, USA*
[c] *Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, 48109, USA*
* *Correspondence to: Matthew Zawistowski, Department of Biostatistics, University of Michigan, Ann Arbor, MI, 48109, USA. E-mail: mattz@umich.edu*

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

that a randomly chosen case has a higher predictive score than a randomly chosen control. Implicit in the prediction modeling and subsequent ROC analysis is the assumption that the observed binary outcomes are measured without error. In practice, this assumption may not hold, meaning samples that are recorded as cases are in truth controls and vice versa. Random error or systematic bias can produce measurement error in a binary variable which is often referred to as misclassification[4]. Examples of this phenomenon are widespread in research, including epidemiology[5, 6, 7], genetics[8, 9], and studies involving administrative claims outcomes[10].

Ignoring misclassification when modeling binary outcomes with logistic regression is well-known to result in biased effect size estimates[4, 11]. Numerous methods have been proposed to obtain bias-corrected, though less efficient, regression parameters estimates, including a modified likelihood equation[11], Bayesian estimation[12], the E-M algorithm[13], simulation[14], and multiple imputation[15]. These papers focused primarily on parameter estimation and did not consider the predictive accuracy of the resulting estimates. Thus, the effect of misclassified outcomes on ROC analysis of risk prediction models has not been well described.

Analysis of misclassified outcomes is receiving renewed interest due to the emergence of research using Electronic Health Record (EHR) data. The promise of precision medicine rests in part on the ability to leverage Big Data resources such as EHRs that track vast clinical, demographic and genetic data into accurate prediction tools for individual patient risk [16, 17, 18]. Already, clinical information recorded in EHRs has been used to develop risk models for a wide range of conditions; examples include cardiovascular disease [19, 20], hospital readmission [21], acute kidney injury [22] and postoperative complications [23]. The clinical research potential of EHRs is immense; however, EHRs are imperfect and highly susceptible to missing and inaccurate diagnoses and behavioral information[24, 25, 26]. In practice, the presence of specific diagnosis or procedural billing codes (e.g. ICD9 codes) in a patient's EHR are often used to identify case samples. Control samples are then defined as patients lacking the codes of interest. There are many reasons that a patient EHR might contain incorrect or missing information, making outcomes extracted from EHRs particularly prone to the problem of misclassification [27]. For example, a code can be erroneously entered in a patient EHR as the result of an incorrect diagnosis, typographical mistake, or ambiguous and heterogeneous use of codes. In addition, because EHRs are rarely shared between hospitals or health systems, whereas patients most certainly are, EHR-based records can be incomplete as patients move between health systems. While statistical methods to create error-adjusted prediction models from misclassified EHR outcomes do exist, the methods to properly assess the prediction accuracy of these models are lacking.

Evaluating diagnostic measurements in the absence of gold standard outcomes is a well-appreciated statistical problem (see [28] for an excellent review). Previous studies have examined ROC analysis for scenarios that include multiple non-gold standard diagnostic tests [29], partial outcome verification leading to missing outcomes [30], prediction based on multiple biomarkers [31] and even no observed outcomes [32]. In this paper, we focus on the specific problem of ROC analysis on risk prediction scores developed from and tested on outcomes subject to misclassification, a scenario commonly encountered in EHR research. Because the risk prediction scores are generated using the misclassified outcomes, they are themselves subject to bias [4, 11, 13, 14]. Even if the prediction model is properly corrected using one of the aforementioned methods, one is left with the dilemma of how to use the imperfect outcomes when performing the ROC analysis.

Here we report that using misclassified outcomes in a standard ROC analysis leads to biased AUC estimates. The AUC bias exists regardless of whether the the predictive model fit on the misclassified outcomes was properly corrected. The extent of the AUC bias depends on multiple factors including the the composition of true cases and controls in the dataset and the specific rates at which each are misclassified. We present an intuitive and computationally simple correction to the standard ROC analysis that accounts for the misclassification inherent in the data. In particular, when computing the coordinates of the ROC curve, we replace the observed and potentially misclassified binary outcomes with a quantitative measurement: the probability that a sample is in truth a case conditional on their predictor covariates, observed outcome and the likelihood that their observed outcome is misclassified. We show, through simulation, that this ROC correction

produces nearly the same AUC value that would have been obtained in the absence of misclassification. Our method builds on the likelihood-based model for correcting logistic regression parameter estimates originally proposed by Neuhaus[11] and likewise assumes that probabilities defining the misclassification mechanism are known or can be estimated. Similar to regression parameter estimation in the presence of misclassification, the AUC estimates suffer from a loss of efficiency compared to AUC based on the true outcomes. However, the AUC bias correction remains effective even in the presence of high rates of misclassification.

Finally, we use the proposed ROC correction to assess a prediction model for inpatient hospitalization among patients in the Veteran's Health Administration (VHA). We extracted clinical and demographic predictors as well as instances of inpatient hospitalization from the VHA EHR for a cohort of over one million samples. However, it is possible that a hospitalization event is not recorded in the VHA EHR if the patient received the care outside of a VHA facility, leading to misclassification of patients with true hospitalizations in our cohort. Ignoring the misclassified hospitalization events results in an underestimate of the true predictive capacity of the model. We show that the misclassification rates can be estimated using an internally validated "gold standard" subset created from Medicare data that captures hospitalization events missing from the VHA EHR and incorporated into our corrected ROC procedure.

3. Methods

3.1. Binary Misclassification Model

We will adapt the basic notation for misclassified binary outcome data used by Neuhaus[11]. Let T be a binary outcome variable and assume that the probability of event T depends on a set of covariate variables $\mathbf{X} = X_1, ..., X_p$ and effect size parameters $\beta = \beta_0, \beta_1, ..., \beta_p$ through the standard logistic model

$$logit[P(T=1|\mathbf{X},\beta)] = log\left(\frac{P(T=1|\mathbf{X})}{1-P(T=1|\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p = \mathbf{X}'\beta. \quad (1)$$

We assume however that the true outcome variable T is not observed. Instead, we observe Y, a realization of the outcome variable subject to misclassification according to the following functions:

$$\gamma_0(\mathbf{X}) = P(Y=1|T=0,\mathbf{X}) \text{ and } \gamma_1(\mathbf{X}) = P(Y=0|T=1,\mathbf{X}). \quad (2)$$

The observed outcome variable Y is then related to the covariates $\mathbf{X}$ through the likelihood equation

$$\begin{aligned} P(Y=1|\mathbf{X},\beta) &= \sum_{t=0}^{1} P(Y=1|\mathbf{X},T) \times P(T|\mathbf{X}) \\ &= \{1-\gamma_1(\mathbf{X})-\gamma_0(\mathbf{X})\} \times P(T=1|\mathbf{X}) + \gamma_0(\mathbf{X}) \\ &= \{1-\gamma_1(\mathbf{X})-\gamma_0(\mathbf{X})\} \times \frac{exp(\mathbf{X}'\beta)}{1+exp(\mathbf{X}'\beta)} + \gamma_0(\mathbf{X}) \end{aligned} \quad (3)$$

We assume that the misclassification parameters satisfy $\gamma_0(\mathbf{X}), \gamma_1(\mathbf{X}) \leq 0.5$, indicating that observed values are no worse than chance. Moreover, the constraint $\gamma_0(\mathbf{X}) + \gamma_1(\mathbf{X}) \leq 1$ is required for numerical estimation of the β values.

3.2. Estimation of Regression Parameters

Assume a dataset consisting of covariate values and binary outcomes is generated using the model described above. We will define three estimates for the effect size parameters β. First, let $\hat{\beta}^T$ be the maximum likelihood estimates for equation (1) had the true observations been observed (i.e. **Y=T**). The $\hat{\beta}^T$ are then the standard log odds ratios obtained from logistic regression. Next, let $\hat{\beta}^I$ denote parameter estimates obtained by ignoring the misclassification and performing a

standard logistic regression of the misclassified outcome Y on the covariates X. That is, $\hat{\beta}^I$ is computed by maximizing the incorrectly specified likelihood function $logit[P(Y=1|\mathbf{X})] = \mathbf{X}'\beta$. The $\hat{\beta}^I$ estimators are biased, with the amount of bias determined by the misclassification functions in equation (2) [11].

Finally, let $\hat{\beta}^M$ be the misclassification-adjusted log odds ratio estimates obtained by maximizing the corrected likelihood function based on equation (3). We assume that the misclassification parameters $\gamma_0(\mathbf{X})$ and $\gamma_0(\mathbf{X})$ are known. The resulting $\hat{\beta}^M$ are consistent estimators of β provided the maximum likelihood regularity conditions are satisfied.

3.3. Prediction Probabilities

After fitting a logistic regression, the effect size estimates can be used in the inverse logit function $\phi(\mathbf{x}, \mathbf{b}) = exp(\mathbf{x}'\mathbf{b})/\big(1 + exp(\mathbf{x}'\mathbf{b})\big)$ to create a prediction model for the binary outcome. Here, we define prediction models based on the three different effect size estimators. Let $\hat{P}^T(X) = \phi(X, \hat{\beta}^T)$ denote the risk prediction model that uses the effect size parameter estimates $\hat{\beta}^T$ obtained when the true outcomes are observed. Likewise, let $\hat{P}^I(X) = \phi(X, \hat{\beta}^I)$ and $\hat{P}^M(X) = \phi(X, \hat{\beta}^M)$ be risk prediction models based on the parameter estimates $\hat{\beta}^I$ and $\hat{\beta}^M$, respectively. In the presence of misclassified outcomes, $\hat{P}^I(X)$ is a naive risk prediction model for the true outcome T based on the biased parameter estimates $\hat{\beta}^I$. Risk scores obtained from this model are in fact prediction values for the misclassified outcome Y rather than the true outcome T. The risk model $\hat{P}^M(X)$ gives prediction probabilities for the true, unobserved outcome by incorporating the bias-corrected parameter estimates $\hat{\beta}^M$.

3.4. ROC Analysis

3.4.1. Standard ROC analysis We will first develop notation for computing the standard Receiver Operating Characteristic (ROC) curve for an arbitrary set of binary outcomes $\mathbf{B} = B_1, B_2, ..., B_N$ and corresponding risk prediction scores $\mathbf{q} = q_1, q_2, ..., q_N$. The ROC curve is defined as the true positive and false positive rates over a range of potential prediction score cutpoints. The coordinates of the ROC curve for outcomes $\mathbf{B}$ and prediction scores $\mathbf{q}$ at a given cutpoint α are defined as $ROC(\alpha, \mathbf{B}, \mathbf{q}) = \Big(FP(\alpha, \mathbf{B}, \mathbf{q}), TP(\alpha, \mathbf{B}, \mathbf{q})\Big)$ where

$$TP(\alpha, \mathbf{B}, \mathbf{q}) = \frac{\sum_{i=1}^{N} I(B_i = 1) \times I(q_i > \alpha)}{\sum_{i=1}^{N} I(B_i = 1)} \tag{4}$$

$$FP(\alpha, \mathbf{B}, \mathbf{q}) = \frac{\sum_{i=1}^{N} I(B_i = 0) \times I(q_i > \alpha)}{\sum_{i=1}^{N} I(B_i = 0)} \tag{5}$$

are the true positive and false positive rates, respectively, and $I()$ is the indicator function. The Area Under the ROC Curve is defined to be $AUC(\mathbf{B}, \mathbf{q}) = \int_\alpha ROC(t, \mathbf{B}, \mathbf{q})dt$ and can be computed numerically using, for example, Riemann Sums or Monte Carlo methods. Using this notation, let $AUC(\mathbf{T}, \hat{P}^T)$ be the AUC for an ROC analysis of the true outcomes T; $AUC(\mathbf{Y}, \hat{P}^I)$ be the AUC value for an ROC analysis of the biased risk scores and the observed outcomes $\mathbf{Y}$; and $AUC(\mathbf{Y}, \hat{P}^M)$ the AUC value for the unbiased risk scores and the observed outcomes $\mathbf{Y}$.

3.4.2. Misclassification Adjusted ROC analysis In the presence of misclassification, there is uncertainty regarding the observed outcomes meaning that cases and controls can be incorrectly tallied in equations (4) and (5). To account for this uncertainty, we propose replacing the observed and potentially misclassified outcome Y in the computation of the true and false positive rates with the "conditional predictive probability" that a sample is in truth a case (i.e. $T = 1$). This predictive probability incorporates information from the observed outcome Y, the bias-corrected prediction probability $\hat{P}^M(X)$ and the misclassification values $\gamma_0(X)$ and $\gamma_1(X)$. As with the misclassification adjusted regression, we assume that $\gamma_0(X)$ and $\gamma_1(X)$ are either known or can be estimated and have previously been used in the modeling building step to obtain the bias-corrected parameter estimates $\hat{\beta}^{\mathbf{M}}$.

The conditional predictive probability that the unobserved outcome is in truth a case is

$$\hat{P}(T=1|Y,X,\hat{\beta}^M) = \frac{P(Y|T=1,X)\times\hat{P}(T=1|X,\hat{\beta}^M)}{\hat{P}(Y|X,\hat{\beta}^M)}$$

$$= \begin{cases} \frac{\left[1-\gamma_1(X)\right]\times\hat{P}^M(X)}{\left[1-\gamma_1(X)-\gamma_0(X)\right]\times\hat{P}^M(X)+\gamma_0(X)} & Y=1 \\ \frac{\gamma_1(X)\times\hat{P}^M(X)}{1-\left[\left[1-\gamma_1(X)-\gamma_0(X)\right]\times\hat{P}^M(X)+\gamma_0(X)\right]} & Y=0 \end{cases}$$

$$= \frac{\left[\gamma_1(X)-Y\times(2\gamma_1(X)-1)\right]\times\hat{P}^M(X)}{(1-Y)+(-1)^{1-Y}\left\{\left[1-\gamma_1(X)-\gamma_0(X)\right]\times\hat{P}^M(X)+\gamma_0(X)\right\}} \quad (6)$$

where the denominator in the first line comes from equation (3). After computing the conditional predictive probability in equation (6) for each sample in the discrimination dataset, we define the misclassification-adjusted ROC curve to be $ROC_M(\alpha) = (FP_M(\alpha), TP_M(\alpha))$ where

$$TP_M(\alpha) = \frac{\sum_{i=1}^N \hat{P}(T_i=1|Y_i,X_i,\hat{\beta}^M)\times I(\hat{P}^M(X_i)>\alpha)}{\sum_{i=1}^N \hat{P}(T_i=1|Y_i,X_i,\hat{\beta}^M)} \quad (7)$$

$$FP_M(\alpha) = \frac{\sum_{i=1}^N \hat{P}(T_i=0|Y_i,X_i,\hat{\beta}^M)\times I(\hat{P}^M(X_i)>\alpha)}{\sum_{i=1}^N \hat{P}(T_i=0|Y_i,X_i,\hat{\beta}^M)} \quad (8)$$

and the corresponding misclassification-adjusted area under the curve, $AUC_M = \int_\alpha ROC_M(t)dt$, can be computed in the typical manner.

4. Simulation Results

We present simulation results that describe the behavior of various strategies for assessing prediction models in the presence of misclassified binary outcomes. We simulated data according to the binary misclassification model described in section 3.1 using R software. For each dataset, we fixed the effect sizes β and misclassification functions $\gamma_0(X)$ and $\gamma_1(X)$. Covariate values X were drawn from standard normal distributions and true binary outcomes T were simulated based on the logistic model in equation (1). We then created an observed version Y of the outcome by changing true cases ($T=1$) to observed controls ($Y=0$) with probability $\gamma_1(X)$, and true controls ($T=0$) to observed cases ($Y=1$) with probability $\gamma_0(X)$. We partitioned each dataset into training and testing cohorts. The bias-corrected regression parameter estimates $\hat{\beta}^M$ were computed on training samples only using an iteratively weighted least squares maximization of the logistic regression based on equation (3) [11]. ROC computations were performed only on the testing samples. Results presented in this paper are for datasets of 5000 training samples and 5000 testing samples. The reported AUC and bias values are averaged over 500 simulated realizations with fixed parameter settings.

For each simulated misclassification dataset, we computed the following four measures of area under the ROC curve:

1. True Outcome ROC Analysis: This scenario assumes the true outcomes are observed ($Y=T$). A logistic regression model is fit on the observed/true outcomes and a standard ROC analysis is performed. For this scenario, we report

$AUC(\mathbf{T}, \hat{P}^T)$ which we treat as the true predictive value when assessing the performance of the next three ROC strategies.

2. Misclassified Outcome ROC Analysis: This scenario assumes that misclassified outcomes are observed, but the misclassification is ignored in both the model fitting and subsequent ROC analysis. Biased parameter estimates $\hat{B}^I$ are used for the prediction model and a standard ROC analysis of the observed outcomes is performed, giving $AUC(\mathbf{Y}, \hat{P}^I)$
3. Corrected Predictions, Standard ROC Analysis: This scenario assumes that misclassified outcomes are observed, and that the misclassification is accounted for in the regression model but ignored in the ROC analysis. Therefore, bias-corrected parameter estimates $\hat{B}^M$ are used for the prediction model but a standard ROC analysis of the observed outcomes is performed, resulting in $AUC(\mathbf{Y}, \hat{P}^M)$.
4. Misclassification Adjusted ROC Analysis: This scenario assumes that misclassified outcomes are observed, and that the misclassification is accounted for in both the regression model and the ROC analysis. That is, the bias-corrected parameter estimates $\hat{B}^M$ are used for the prediction model, conditional predictive probabilities (eq. 6) are computed and incorporated into the ROC analysis, leading to AUC_M.

The goal of the misclassification adjusted ROC analysis for a given dataset is to reproduce the AUC value that would have been observed had there been no misclassification of outcomes. Therefore we define AUC bias for the naive or corrected ROC analyses above as the difference between the AUC from true outcome ROC analysis, $AUC(\mathbf{T}, \hat{P}^T)$, and the AUC from the respective strategies ($AUC(\mathbf{Y}, \hat{P}^I)$, $AUC(\mathbf{Y}, \hat{P}^M)$ or AUC_M). In the following, we report the mean value of bias across realizations of datasets from the same underlying model parameters.

[Table 1 about here.]

4.1. Bias in AUC of Misclassified Outcomes

We first show the results of standard ROC analysis performed on true and misclassified versions of simulated binary outcome data (scenarios 1 and 2 above) to establish the existence of AUC bias and to determine conditions that are most problematic. Table 1 gives the mean AUC values from an ROC analysis of true outcomes $AUC(\mathbf{T}, \hat{P})$ and from a naive ROC analysis of the misclassified outcomes $AUC(\mathbf{Y}, \hat{P}^I)$. To allow insight, we simulated outcomes with only a single covariate and constant misclassification at combinations of realistically low and high parameter values for the outcome-covariate model (β_0 and β_1) and the misclassification model (γ_0 and γ_1).

Misclassification leads to bias in the estimate of AUC for all parameter combinations. When the effect size β_1 is small, the true AUC value is already relatively low and any misclassification introduces only a small bias in AUC. When the effect size is larger ($\beta_1 = 1.0$), misclassification produces a greater bias in AUC, particularly for very rare events (baseline risk of 0.01). For example, misclassifying rare events, even at low false positive (γ_0) and false negative (γ_1) rates of 5%, leads to a 26% reduction in AUC value, from 0.758 down to 0.560. Misclassifying true controls as cases has a more dramatic effect on AUC bias. Whereas increasing the false negative rate to $\gamma_1 = 20\%$ in the prior scenario leads to only a slight increase in AUC bias (from 26% to a 27.2% reduction), increasing the false positive rate to $\gamma_0 = 20\%$ results in an AUC value that is reduced by 31.6% from the true value.

The increased sensitivity of AUC bias to false positive misclassification rate (γ_0) in the above simulations stems from the fact that controls far outnumber cases in those datasets, as often occurs in population cohort studies. Since the vast majority of samples are true controls, the false positive rate has a larger influence on the actual number of outcomes that will be misclassified. We repeated this analysis using β_0 values in which cases are more prevalent than controls (Supplementary Table 2). As expected, in that scenario, it is the false negative misclassification rate (γ_1) that has the greater effect on AUC bias. Thus, the prevalence of the outcome in the dataset is critical in determining which type of misclassification will have the greater effect on AUC bias.

[Figure 1 about here.]

4.2. AUC Correction

4.2.1. Constant Misclassification Initially, we assume the misclassification functions $\gamma_0(X)$ and $\gamma_1(X)$ are known and constant across all samples. Figure 1 shows the distribution of AUC values for the four ROC procedures based on datasets simulated with effect sizes of $\beta_0 = -1, \beta_1 = 1$ at reasonably large constant misclassification rates of $\gamma_0 = 0.2 and \gamma_1 = 0.3$.

The first boxplot shows the distribution of $AUC(\mathbf{T}, \hat{P})$, the AUC values that would be obtained if the true outcomes were observed, and therefore provides the true distribution that we wish to recover. The dotted red line at 0.741 marks the mean $AUC(\mathbf{T}, \hat{P})$ value for the True Outcome ROC analysis. The second boxplot shows the distribution of $AUC(\mathbf{Y}, \hat{P}^I)$, an analysis that ignores misclassification in both the model fitting and ROC procedure. As expected, this naive analysis produces biased, underestimates of true AUC (mean bias=0.129). The third boxplot shows values of $AUC(\mathbf{Y}, \hat{P}^M)$ based on a standard ROC analysis of misclassification-corrected prediction estimates. Interestingly, these AUC values are virtually identical to those from the previous ROC analysis that completely ignored the misclassification. Therefore, accounting for misclassification in the modeling step had little effect on AUC computation when the misclassified outcomes were still used in a standard ROC analysis.

Finally, the fourth boxplot shows AUC_M, the AUC values based on our misclassification adjusted ROC method. The distribution for AUC_M is centered over the mean of the true AUC values (dotted red line) and the ROC adjustment has removed nearly all bias introduced by the misclassification (mean bias=-0.001). The AUC_M estimates do, however, have a larger variance than the $AUC(\mathbf{T}, \hat{P})$ from the true model. This is expected since the bias-corrected parameters estimates $\hat{\beta}^M$ are themselves known be less efficient than estimates based on the true data $\hat{\beta}^T$ [11].

We computed confidence intervals for the corrected AUC_M estimator using a bootstrap technique. We created bootstrapped training datasets by randomly drawing samples with replacement from the the training cohort only, recomputing the $\hat{\beta}^M$ and prediction model for each bootstrapped set and calculating the corresponding AUC_M in the original testing data. Assuming that AUC_M is an estimator for $AUC(\mathbf{T}, \hat{P})$, the true AUC value that would have been observed in the absence of misclassified outcomes, we find that these bootstrap-based confidence intervals have accurate coverage probabilities when the underlying misclassification rates are known. For example, 90% confidence intervals for the AUC_M values shown in Figure 1 (based on 300 bootstraps of the training data), contained the true AUC in the testing data for 89% of the 500 simulated realizations.

[Figure 2 about here.]

Figure 2 shows mean bias and associated standard errors for AUC values obtained using a standard ROC analysis of misclassified outcomes and the misclassification-adjusted ROC analysis across a range of misclassification rates (also Supplementary Figure 5). As expected, bias in the standard AUC computation increases with increasing levels of misclassification. The adjusted ROC procedure removes nearly all bias in AUC, although the standard error on these corrected estimates does increase with increasing misclassification. We observed similar results for simulations with multiple predictor variables (Supplementary Figure 6).

[Figure 3 about here.]

4.2.2. Differential Misclassification Next, we relaxed the assumption of constant misclassification, allowing the misclassification rates to be functions of the covariate and thus vary between samples. Figure 3 shows distributions of AUC values for two extreme scenarios of covariate-dependent differential misclassification. Again, we set $\beta_0 = -1$ and $\beta_1 = 1$ and modeled misclassification using logistic functions of the covariate X as follows: $logit\big[\gamma_0(X)\big] = \theta_0 + \theta_1 X$ and $logit\big[\gamma_1(X)\big] = \sigma_0 + \sigma_1 X$. We fixed $\theta_0 = \sigma_0 = -1.386294$ to give $\gamma_0(0) = \gamma_1(0) = 0.2$. Then, in the first scenario, we set $\theta_1 = -0.5$ and $\sigma_1 = 1.5$ so that false positive rate increases with the probability of being a control and false negative rate increases with the probability of being a case (Supplementary Figure 7). AUC values from the misclassified outcome ROC analysis severely underestimate of the true AUC (mean bias=0.179). The AUC_M values from the adjusted ROC

analysis have a mean bias of only -0.001, indicating that the correction method has removed virtually all bias induced by misclassification.

In the second scenario, we reversed the relationship between misclassification rates and event probability. We set $\theta_1 = 1.5$ and $\sigma_1 = -0.5$ so that the false positive rate increases with the probability of being a case and the false negative rate increases with the probability of being a control (Supplementary Figure 8). Here, the misclassified outcome ROC analysis actually overestimates the true AUC value (mean bias=-0.054). Again, the misclassification-adjusted ROC procedure yielded nearly unbiased AUC estimates (mean bias=-0.001).

This differential misclassification example highlights the fact that AUC bias can actually occur in both directions. In the first scenario, the noise introduced by the misclassification weakened the regression association between the covariate and the outcome which ultimately led to a lower estimated predictive capacity. Underestimation of the true AUC is the more frequent scenario, also occurring for constant misclassification as well as when $\gamma_0(X)$ and $\gamma_1(X)$ have the same direction of effect (Supplementary Figure 9). In the second scenario, the misclassification systematically strengthened the perceived association between the covariate and the outcome leading to an inflated estimate of the predictive ability of the covariate. In each case, however, the ability to recover bias-corrected parameter estimates in the modeling step allows the misclassification-adjusted ROC procedure to substantially reduce the AUC bias.

5. Application to EHR Hospitalization Data

We present a proof of principle example using data from the Veteran's Health Administration (VHA) Electronic Health Record (EHR) and Centers for Medicare and Medicaid Services (CMS). This example serves as an application of our proposed ROC correction on real EHR data as well as a description of how the misclassification functions can be estimated using an internally validated "gold standard" subset. The binary outcome of interest is the occurrence of an inpatient hospitalization in VHA patients aged 65 or older during a three year followup period from 1/1/2007 through 12/31/2009. We wish to know how well hospitalization events can be predicted using demographic (age, sex), behaviorial (smoking), medication (hypertension prescription) and comorbidity (diabetes, chronic heart failure, chronic obstructive pulmonary disease, atrial fibrillation, and depression) information collected at baseline. We defined a nationwide cohort of $N = 1,037,428$ VHA healthcare patients between the ages of 65 and 80 at the start of 2006 and with at least two outpatient appointments within the VHA system during 2006. We partitioned the dataset into a training cohort of $750,000$ samples for model building and the remaining $\sim 250,000$ samples to a testing cohort for the ROC analysis. Full description of the cohort and specific diagnosis and procedural codes used for comorbidities are given in the appendix.

5.1. Misclassification of Hospitalization Events

Patients in our cohort may be eligible for medical care through non-VHA sources, meaning that hospitalizations either occurring outside of VHA facilities or not billed to the VHA may not be recorded in the VA EHR. This inevitable gap in recording leads to the clear potential to misclassify patient outcomes when using only VHA EHR data, a common problem for any health care system. We therefore extracted Medicare records from CMS for patients in our cohort to identify additional hospitalization events not appearing in the VHA EHR. We treated the combined set of hospitalization events from the VHA EHR and CMS as the "true" outcomes (T) and pose the question of how using only "observed" VHA EHR outcomes (Y) would affect our prediction modeling. Under this scenario, patients with a hospitalization event in CMS but not in their VHA EHR would be misclassified as non-events.

We purposefully chose inpatient hospitalizations in Medicare-eligible patients ≥ 65 years of age to create an example in which total event capture should be nearly complete between the VA EHR and CMS. Inpatient hospitalizations are typically medically intensive and expensive, leading to a large number of procedural and billing codes that leave a substantial trail in the patient medical record, be it the VHA EHR or CMS, which is unlikely to be missed.

In total, 130,876 patients (12.6%) had an inpatient hospitalization during the three year followup recorded in their VHA EHR. In comparison, 320,697 patients (30.9%) had a hospitalization event based on the combined VHA and CMS records, indicating the extent of false negative misclassification for hospitalization events. We make the simplifying assumption that all hospitalization events are real. Although this may not hold in practice, the number of false positive events is likely much smaller than false negative events. Then, by construction, the false positive misclassification function is $\gamma_0(X) = 0$, and the false negative function $\gamma_1(X)$ is non-zero and unknown but there is evidence that it is covariate dependent (Supplementary Figure 10).

5.2. *Standard ROC Analysis*

Figure 4 shows ROC curves for predicting three year hospitalization events. First, we used the true (VHA+CMS) outcomes in the regression model with standard ROC analysis (black curve) to establish the "true" AUC value in the absence of misclassification. This analysis yielded an AUC of 0.669 in the testing cohort. Next, we determined the effect of ignoring misclassification by fitting the logistic prediction model on the observed (VHA-only) outcomes and using the subsequent biased prediction values in a standard ROC analysis with the observed outcomes (red line). The AUC for this analysis was 0.592. As expected, the missed hospitalizations in the VHA-only outcome led to a biased AUC that lowered the perceived predictive value of the covariates. The 11.5% drop in AUC is consistent with simulation results in Table 1 for fairly common outcomes with low false positive and high false negative rates.

[Figure 4 about here.]

5.3. *Estimation of Misclassification Rates*

In order to apply our ROC procedure we require an estimate of the misclassification function $\gamma_1(X)$. To do this, we randomly selected a subset of the training samples to serve as internally validated "gold standard" outcomes. That is, for these samples, we revealed the true (VHA+CMS) hospitalization outcome and fit the following logistic regression model on true hospitalization cases:

$$logit[P(Y = 1|T = 1, \mathbf{X})] = logit[1 - \gamma_1(X)]. \quad (9)$$

where T is the VHA+CMS outcome, Y is the VHA-only outcome and $\mathbf{X}$ is the same set of covariates used for predicting hospitalizations. We used the regression parameter estimates from equation (9) to compute a covariate-dependent estimate of false negative misclassification $\hat{\gamma_1}(X)$ for each sample in the dataset. Using the $\hat{\gamma_1}(X)$ values and the VHA-only hospitalization outcomes Y, we computed AUC_M using our misclassification adjusted ROC procedure.

5.4. *Misclassification-Adjusted ROC Analysis*

Due to the sampling variation inherent in drawing an internal validation cohort, we created 20 realizations of $5,000$ randomly selected gold standard samples. The gray curves in Figure 4 show the misclassification-adjusted ROC curves based on the different realizations of the internal validation cohorts. The true ROC curve (black) is roughly centered within the distribution of misclassification-adjusted ROC curves. The AUC_M values ranged from 0.626 to 0.692, with an average value of 0.658, providing a much closer estimate of the AUC value of 0.669 from the true outcome ROC analysis.

Coverage of the bootstrap-based confidence intervals suffered (only 2 of the 20 90% confidence intervals covered the true AUC) due to the misclassification rates being estimated rather than known exactly (Supplementary Figure 11). However, none of the confidence intervals covered or were below the naive AUC estimate of 0.592, only two intervals overestimated the true AUC, and, in many cases, the distance between the bounds of the confidence intervals and the true AUC were negligible (e.g. < 0.01). Thus, the combination of a small internally validated subset ($< 0.5\%$ of the full cohort) and our proposed ROC correction were able to sufficiently estimate unknown EHR misclassification rates and provide a much improved estimate of the predictive accuracy for hospitalization events.

6. Discussion

Building accurate predictive models based on EHR data is a critical step toward the goal of personalized medicine. Analysis of misclassified outcome data is therefore only likely to increase as EHR research becomes more prevalent. Properly accounting for inevitable misclassification in Big Data resources is required in both the statistical model building step as well as for determining prediction accuracy. Ignoring outcome misclassification when fitting a regression model leads to biased parameter estimates. Numerous methods have been proposed to correct that bias. Here we have shown that simply correcting the effect estimates in the regression model is not sufficient for performing a valid ROC analysis. Using the misclassified outcomes, even with accurate risk prediction scores, leads to biased estimates of AUC. We have introduced a correction procedure in which the ROC curve incorporates quantitative likelihoods for the outcome into the definitions of true and false positive rates in place of the observed potentially misclassified outcome (equations 7 and 8). This adjustment leads to more accurate estimates of true and false positive rates for any fixed cutoff α which in turn substantially reduces the bias in area under the ROC curve.

A major advantage of our correction method is that it is computationally simple, requiring only one additional value (equation 6) be computed beyond the standard ROC analysis. Computational efficiency is essential in the era of Big Data in which datasets can be on the order of thousands of variables in millions of samples, thereby placing a premium on algorithms that can obtain the desired result without iterating through the data many times or requiring repeated draws of random variables. While the AUC bias could potentially be corrected using simulation, imputation or more complex Bayesian methods, our closed form equation is substantially faster. It is therefore ideal for quickly and accurately comparing the predictive value of competing models in very large datasets or determining the predictive value of individual model covariates.

Analyzing misclassified data requires some knowledge of the underlying misclassification mechanism. Unfortunately this mechanism can be very complicated and unintuitive. Nevertheless, we have assumed that misclassification probabilities are either known or can be estimated. Since the misclassification probabilities are typically needed for correction in the model building step, we do not require additional information for our ROC correction. Estimates of misclassification rates can come from external diagnostic data or potentially even inferred from summary analyses of the observed data. In this paper, we showed how covariate-dependent misclassification probabilities can be estimated from an internally validated subset in which the true outcome is known. We used Medicare records to obtain the true outcome for hospitalization events but other applications may require validation techniques such as medical chart review or molecular-based diagnosis (e.g. biopsy). The feasibility of this strategy is dependent on the cost (both financial and time) of obtaining true validated outcomes. It is important to note that although the actual causal mechanism leading to the misclassification in our data example remained unclear, the sample-level estimates of misclassification were sufficiently accurate to improve AUC estimation.

In practice, it may be unclear how extensive the misclassification is or how accurately it has been estimated. Our results in Tables 1 and 2 provide a means for determining how large the misclassification rates must be for different scenarios in order to have a major impact on AUC estimates. In some cases the misclassification may be judged low enough that the effect on AUC is negligible. When fitting models and performing ROC analysis on misclassified outcomes, we recommend a sensitivity analysis in which AUC is computed for a range of misclassification parameter values to understand how changes in the misclassification probabilities affect inference. Similar sensitivity analyses have already been recommended for misclassification in the model building step[33].

There are many analytic considerations in prediction modeling and discrimination analysis. Throughout we have assumed that the misclassification rates are the same in the testing and training cohorts, as expected when a single dataset is partitioned, however this is not required. Provided the misclassification rates from the training data are used for the model building step and the misclassification rates from the testing data are used in the misclassification-adjusted ROC analysis, the correction method remains valid. We have also used logistic regression throughout as our prediction model, but the prediction scores used in the ROC analysis can come from any type of classification model [2]. We simply assume that the

risk scores have been properly corrected for the misclassification. Finally, sample size plays an important role in analysis of misclassified outcome data. Of course, larger sample sizes improve precision of effect size parameter estimates. But, more importantly, small sample sizes can lead to difficulty in convergence when trying to obtain maximum likelihood estimates for the misclassification-adjusted effect sizes $\hat{\beta}^M$. Here we used the Iteratively Weighted Least Squares (IWLS) approach to solve for the $\hat{\beta}^M$ in training cohorts of 5000 samples. Smaller sample sizes may require alternative maximization procedures.

7. Funding

This work was supported by VA IIR 11-088. Support for VA/CMS data is provided by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Health Services Research and Development, VA Information Resource Center (Project Numbers SDR 02-237 and 98-004).

8. Acknowledgments

We thank Tom Braun, Sebastian Zöllner and Phillip Boonstra for providing valuable feedback on the initial manuscript and the three reviewers for constructive criticisms that improved the manuscript.

References

[1] Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer: New York, 2009.

[2] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer: New York, 2001.

[3] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 1982; **143**(1):29–36, doi:10.1148/radiology.143.1.7063747. URL http://dx.doi.org/10.1148/radiology.143.1.7063747, pMID: 7063747.

[4] Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology* 1977; **105**(5):488–495. URL http://aje.oxfordjournals.org/content/105/5/488.abstract.

[5] Chen Q, Galfalvy H, Duan N. Effects of disease misclassification on exposure-disease association. *American Journal of Public Health* 2013; **103**(5):e67e73, doi:http://doi.org/10.2105/AJPH.2012.300995.

[6] Edwards JK, Cole SR, Chu H, Olshan AF, Richardson DB. Accounting for outcome misclassification in estimates of the effect of occupational asbestos exposure on lung cancer death. *American Journal of Epidemiology* 2014; **179**(5):641–647, doi:http://doi.org/10.1093/aje/kwt309.

[7] Porter KA, Burch CL, Poole C, Juliano JJ, Cole SR, Meshnick SR. Uncertain outcomes: Adjusting for misclassification in antimalarial efficacy studies. *Epidemiology and Infection* 2011; **139**(4):544–551, doi:http://doi.org/10.1017/S0950268810001652.

[8] Colhoun HM, McKeigue PM, Smith GD. Problems of reporting genetic associations with complex outcomes. *The Lancet* 2003; **361**(9360):865 – 872, doi:http://dx.doi.org/10.1016/S0140-6736(03)12715-8. URL http://www.sciencedirect.com/science/article/pii/S0140673603127158.

[9] Smith S, Hay EH, Farhat N, Rekaya R. Genome wide association studies in presence of misclassified binary responses. *BMC Genetics* 2013; **14**:124, doi:http://doi.org/10.1186/1471-2156-14-124.

[10] Funk MJ, Landi SN. Misclassification in administrative claims data: Quantifying the impact on treatment effect estimates. *Current epidemiology reports* 2014; **1**(4):175–185.

[11] Neuhaus J. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 1999; **86**(4):843–855, doi:10.1093/biomet/86.4.843. URL http://biomet.oxfordjournals.org/content/86/4/843.abstract.

[12] McInturff P, Johnson WO, Cowling D, Gardner IA. Modelling risk when binary outcomes are subject to error. *Statistics in Medicine* 2004; **23**(7):1095–1109, doi:10.1002/sim.1656. URL http://dx.doi.org/10.1002/sim.1656.

[13] Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* 1997; **146**(2):195–203. URL http://aje.oxfordjournals.org/content/146/2/195.abstract.

[14] Kuchenhoff H, Mwalili SM, Lesaffre E. A general method for dealing with misclassification in regression: The misclassification simex. *Biometrics* 2006; **62**(1):85–96, doi:10.1111/j.1541-0420.2005.00396.x. URL http://dx.doi.org/10.1111/j.1541-0420.2005.00396.x.

[15] Edwards JK, Cole SR, Troester MA, Richardson DB. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *American Journal of Epidemiology* 2013; doi:10.1093/aje/kws340. URL http://aje.oxfordjournals.org/content/early/2013/04/03/aje.kws340.abstract.

[16] Collins FS, Varmus H. A new initiative on precision medicine. *New England Journal of Medicine* 2015; **372**(9):793–795, doi:10.1056/NEJMp1500523. URL http://dx.doi.org/10.1056/NEJMp1500523, pMID: 25635347.

[17] Jameson JL, Longo DL. Precision medicine : Personalized, problematic, and promising. *New England Journal of Medicine* 2015; **372**(23):2229–2234, doi:10.1056/NEJMsb1503104. URL http://dx.doi.org/10.1056/NEJMsb1503104, pMID: 26014593.

[18] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* 2012; **13**(6):395–405.

[19] Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Medical Care* 2013; **51**(3):251–258, doi:10.1097/MLR.0b013e31827da594.

[20] Pike MM, Decker PA, Larson NB, St Sauver JL, Takahashi PY, Roger VL, Rocca WA, Miller VM, Olson JE, Pathak J, *et al.*. Improvement in cardiovascular risk prediction with electronic health records. *Journal of Cardiovascular Translational Research* 2016; :1–9doi:10.1007/s12265-016-9687-z. URL http://dx.doi.org/10.1007/s12265-016-9687-z.

[21] Nguyen OK, Makam AN, Clark C, Zhang S, Xie B, Velasco F, Amarasingham R, Halm EA. Predicting all-cause readmissions using electronic health record data from the entire hospitalization: Model development and comparison. *Journal of Hospital Medicine* 2016; :n/a–n/adoi:10.1002/jhm.2568. URL http://dx.doi.org/10.1002/jhm.2568.

[22] Matheny ME, Miller RA, Ikizler TA, Waitman LR, Denny JC, Schildcrout JS, Dittus RS, Peterson JF. Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. *Medical Decision Making* 2010; doi:10.1177/0272989X10364246. URL http://mdm.sagepub.com/content/early/2010/03/30/0272989X10364246.abstract.

[23] Soguero-Ruiz C, Hindberg K, Mora-Jimnez I, Rojo-lvarez JL, Skrvseth SO, Godtliebsen F, Mortensen K, Revhaug A, Lindsetmo RO, Augestad KM, *et al.*. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *Journal of Biomedical Informatics* 2016; **61**:87–96, doi:http://dx.doi.org/10.1016/j.jbi.2016.03.008. URL http://www.sciencedirect.com/science/article/pii/S1532046416000484.

[24] Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Medicine* 2015; **7**(1):1–14, doi:10.1186/s13073-015-0166-y. URL http://dx.doi.org/10.1186/s13073-015-0166-y.

[25] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association* 2013; **20**(1):144–151, doi:10.1136/amiajnl-2011-000681. URL http://jamia.oxfordjournals.org/content/20/1/144.

[26] Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* 2013; **20**(1):117–121, doi:10.1136/amiajnl-2012-001145. URL http://jamia.oxfordjournals.org/content/20/1/117.

[27] Wells B, Chagin K, Nowacki A, Kattan M. Strategies for handling missing data in electronic health record derived data. *eGEMs* 2013; **1**(3):1035, doi:http://doi.org/10.13063/2327-9214.1035.

[28] Collins J, Huynh M. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statistics in Medicine* 2014; **33**(24):4141–4169. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4199084/?report=classic.

[29] Jones G, Johnson WO, Hanson TE, Christensen R. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics* 2010; **66**(3):855–863, doi:10.1111/j.1541-0420.2009.01330.x. URL http://dx.doi.org/10.1111/j.1541-0420.2009.01330.x.

[30] Pepe MS, Alonzo TA. Comparing disease screening tests when true disease status is ascertained only for screen positives. *Biostatistics* 2001; **2**(3):249–260, doi:10.1093/biostatistics/2.3.249. URL http://biostatistics.oxfordjournals.org/content/2/3/249.abstract.

[31] Jafarzadeh SR, Johnson WO, Gardner IA. Bayesian modeling and inference for diagnostic accuracy and probability of disease based on multiple diagnostic biomarkers with and without a perfect reference standard. *Statistics in Medicine* 2016; **35**(6):859–876, doi:10.1002/sim.6745. URL http://dx.doi.org/10.1002/sim.6745, sIM-13-0575.R2.

[32] Branscum AJ, Johnson WO, Hanson TE, Baron AT. Flexible regression models for roc and risk analysis, with or without a gold standard. *Statistics in Medicine* 2015; **34**(30):3997–4015, doi:10.1002/sim.6610. URL http://dx.doi.org/10.1002/sim.6610, sim.6610.

[33] Lyles RH, Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Statistics in Medicine* 2010; **29**(22):2297–2309.

9. Supplementary Material

[Table 2 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

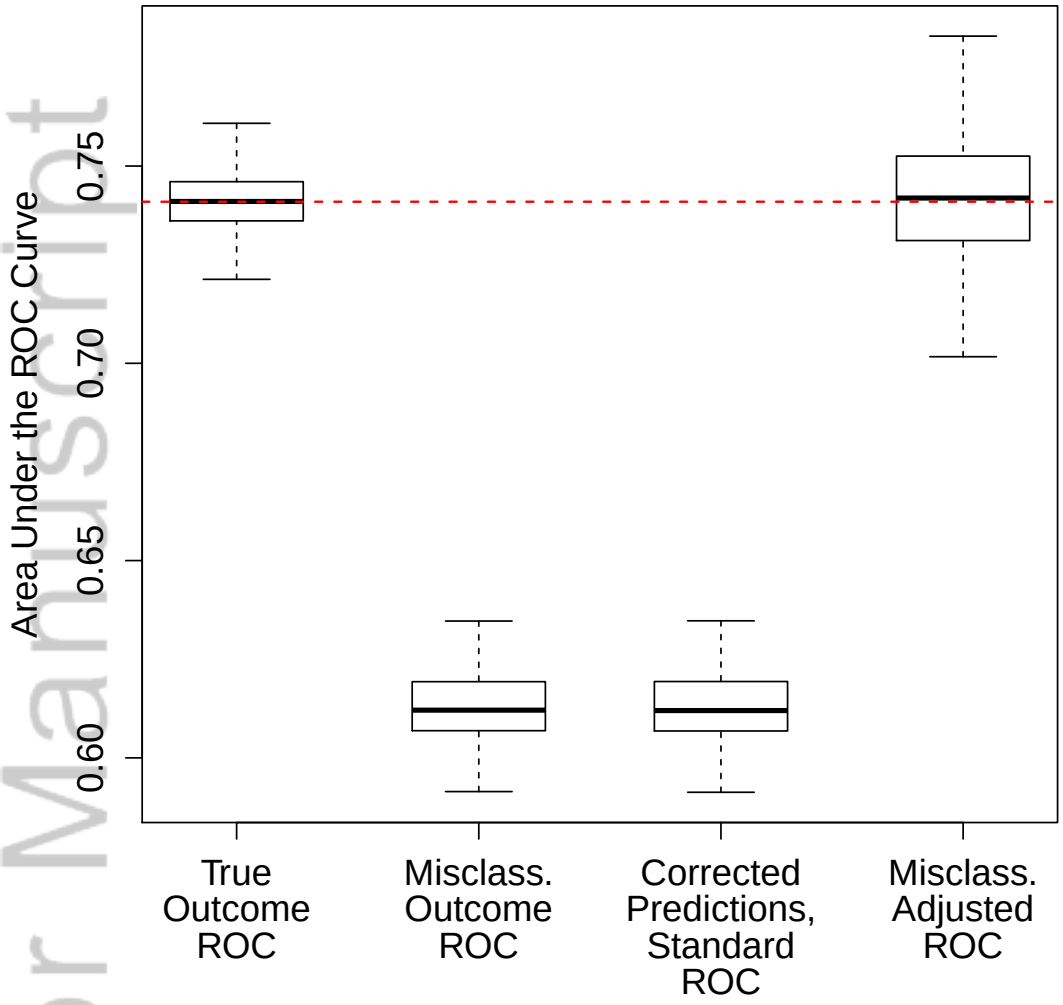

Figure 1. Distribution of AUC values for four strategies of performing ROC analysis in the presence of misclassified binary outcomes. (1) The distribution of AUC values when the true outcomes are observed. (2) Ignoring misclassification in both the regression prediction model and ROC analysis leads to a biased, underestimate of the true AUC. (3) The AUC bias remains when accounting for the misclassified outcomes during the modeling step, but ignoring them in ROC computation. (4) The misclassification-adjusted ROC procedure gives nearly unbiased AUC estimates (bias=-0.001) but larger variance than the true outcome AUC. Results are shown for 500 simulated datasets of 5000 training and 5000 testing samples with model effect sizes of $\beta_0 = -1, \beta_1 = 1$ and constant misclassification parameters $\gamma_0 = 0.2, \gamma_1 = 0.3$.

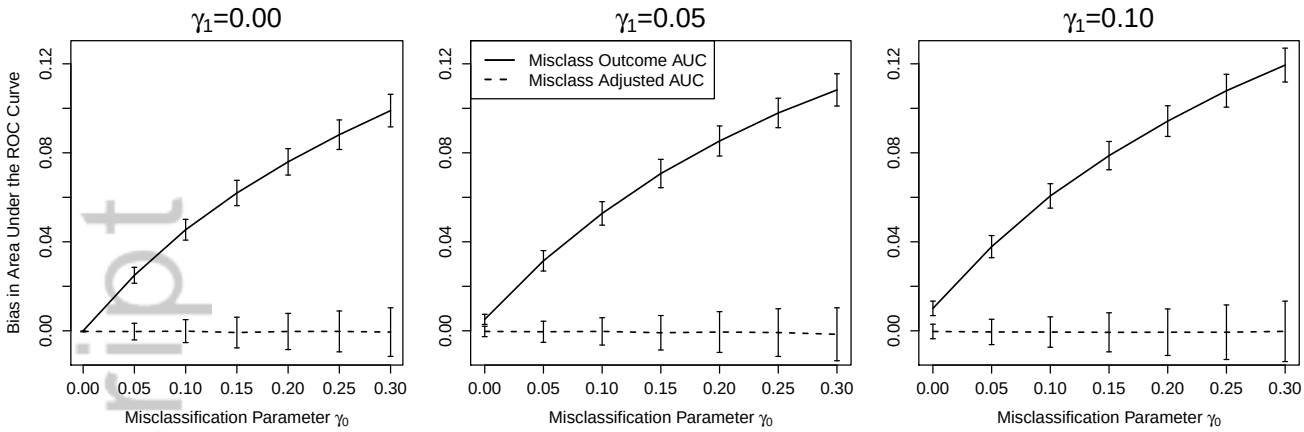

Figure 2. Bias in AUC for standard and misclassification-adjusted ROC analysis over a range of constant misclassification rates ($\beta_0 = -1, \beta_1 = 1$). Mean bias in AUC for a standard ROC analysis of misclassified outcomes (solid lines) increases with increasing amount of misclassification. The misclassification-adjusted ROC procedure (dashed lines) has bias of nearly zero over all combinations of false positive and false negative misclassification rates. Standard error on the misclassification-adjusted AUC estimates (vertical bars) increases with increasing rates of misclassification.

Statistics
in Medicine

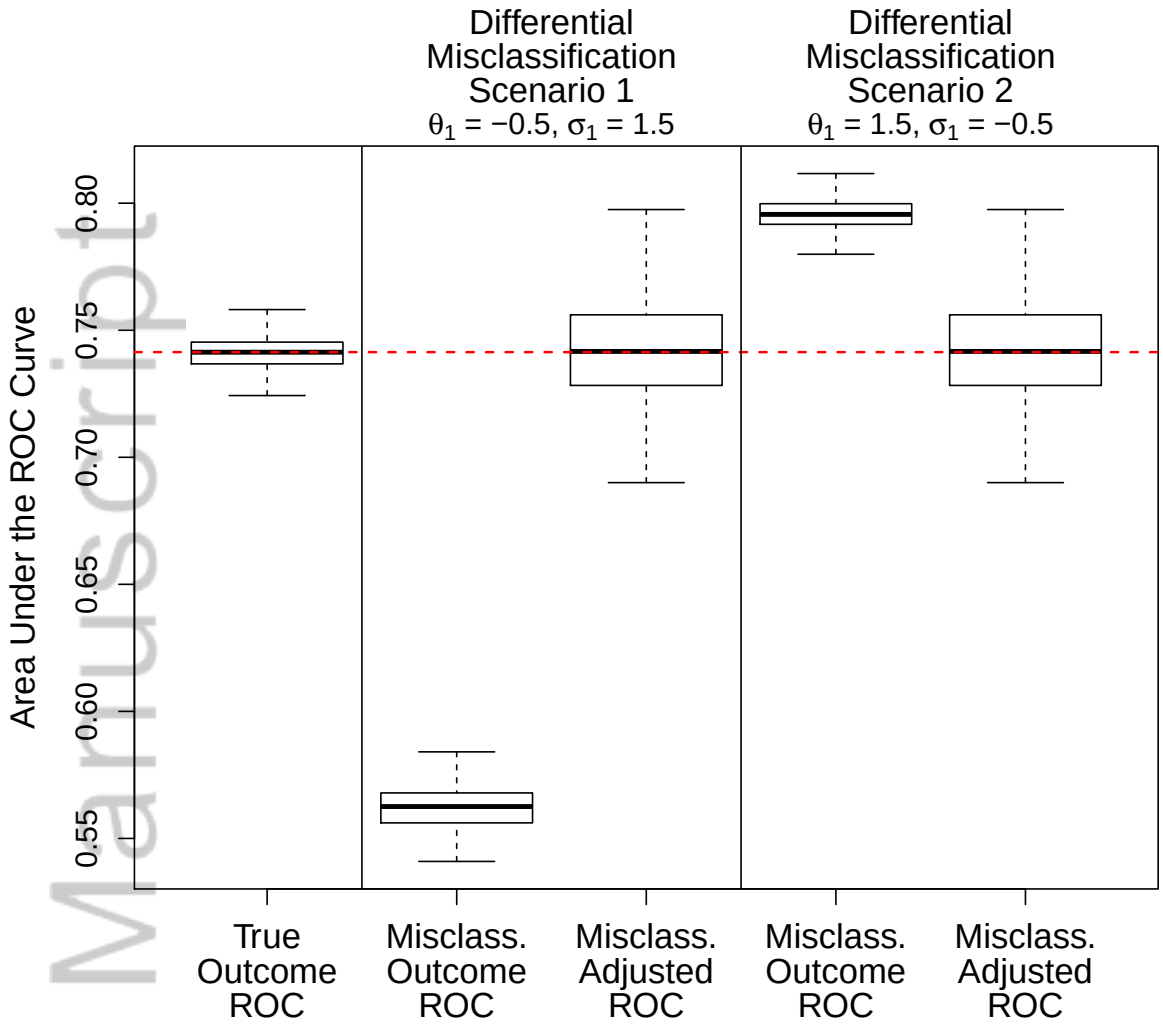

Figure 3. Distribution of AUC values for covariate-dependent differential misclassification. In the first scenario, the differential misclassification reduces the true association between covariate and outcome, leading to an underestimate of the true AUC. In the second scenario, the misclassification rates inflate the covariate-outcome association in the prediction model, and lead to an overestimate of the true AUC. In both cases, the misclassification-adjusted ROC procedure removes nearly all bias in AUC estimates.

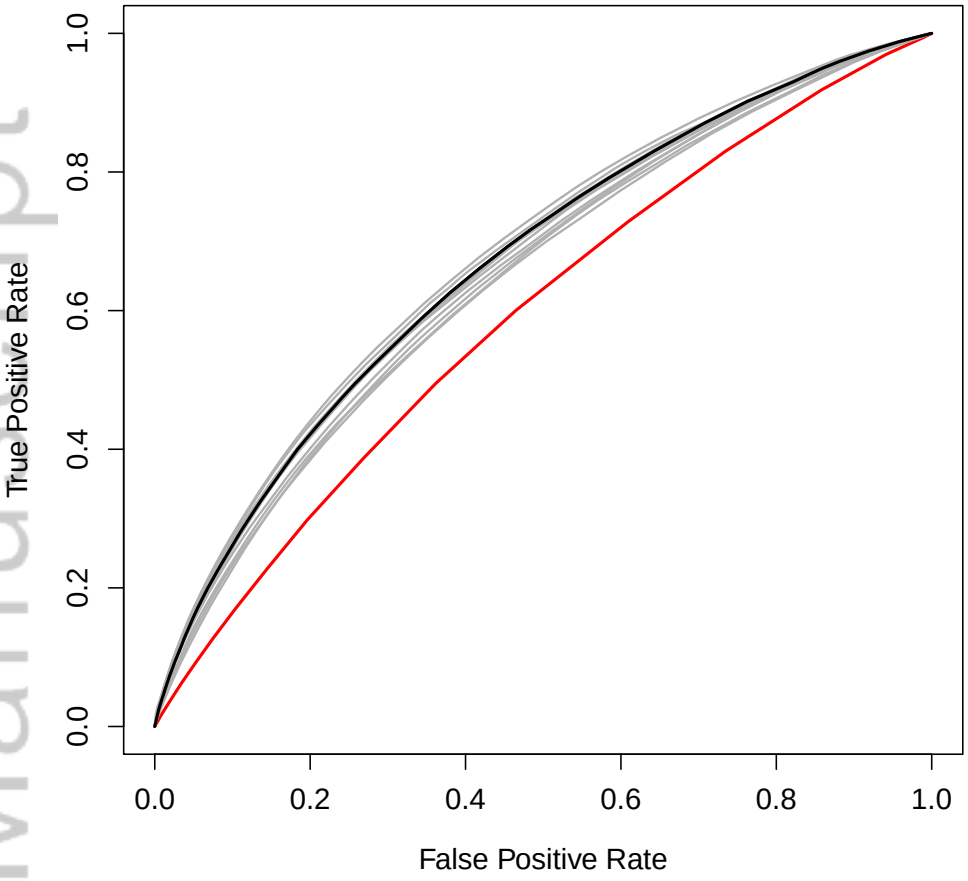

Figure 4. ROC curves for prediction of hospitalization events in Veteran's Health Administration (VHA) patients. The ROC curve (black) obtained when the true outcomes are observed has AUC=0.669. Using only hospitalization events recorded in the VHA EHR leads to misclassified outcomes. A standard ROC analysis of only VHA outcomes (red curve) has AUC=0.592, an underestimate of the true predictive capability. We applied the misclassification-adjusted ROC procedure to the data using "gold standard" subsets with validated outcomes to model outcome misclassification in the VHA-only data. The gray lines show the misclassification-adjusted ROC curve for realizations of the gold standard subset. The mean AUC value from the misclassification-adjusted ROC curves was 0.658, meaning the combination of a small gold standard subset and adjusted ROC procedure dramatically improved AUC estimation.

Statistics
in Medicine

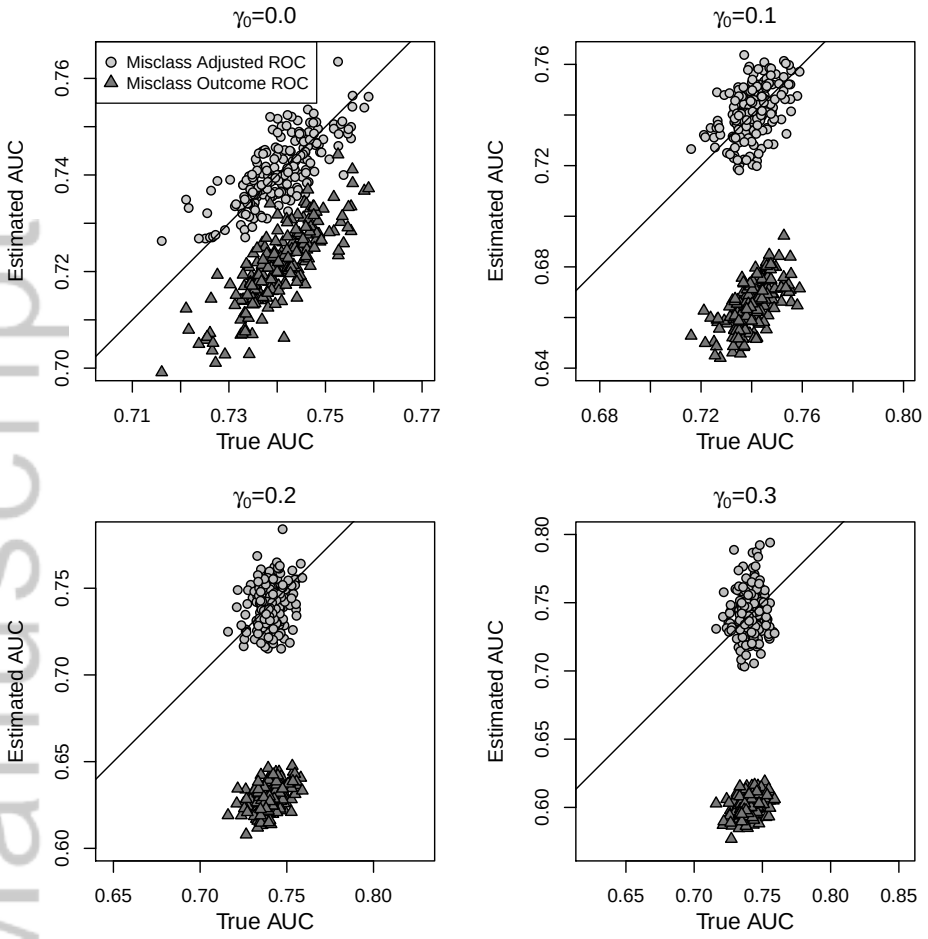

Figure 5. Scatterplots of standard AUC estimates and misclassification-adjusted AUC estimates versus the true AUC values for various levels of misclassification. In each figure, simulations were for model parameters $\beta_0 = -1$ and $\beta_1 = 1$ and fixed false-negative misclassification rate of $\gamma_1 = 0.2$. As misclassification increases, bias increases in standard estimates of AUC (triangles). The misclassification-adjusted AUC estimators (circles) have bias of nearly zero but increasing variance as misclassification increases.

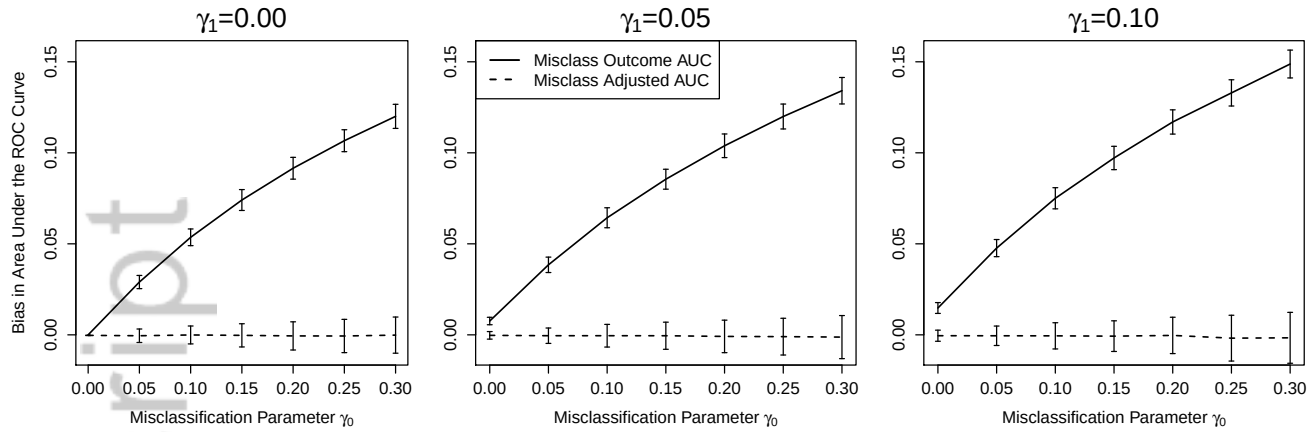

Figure 6. Bias in AUC for standard and misclassification-adjusted ROC procedures over a range of constant misclassification values for a model with multiple predictor variables ($\beta_0 = -1, \beta_1 = 1, \beta_2 = 0.5, \beta_3 = -1$). For each simulated dataset, we computed bias as the difference between the AUC value based on the true outcome data $AUC(\mathbf{T}, \hat{P})$ and the estimates using the observed outcome data $AUC(\mathbf{Y}, \hat{P}^M)$ and AUC_M. Mean bias in AUC for standard ROC computation (solid lines) increases with increasing amount of misclassification. The proposed misclassification-adjusted ROC procedure (dashed lines) has bias of nearly zero over all combinations of false positive and false negative values. The vertical bars give standard errors for the AUC estimates.

Statistics
in Medicine

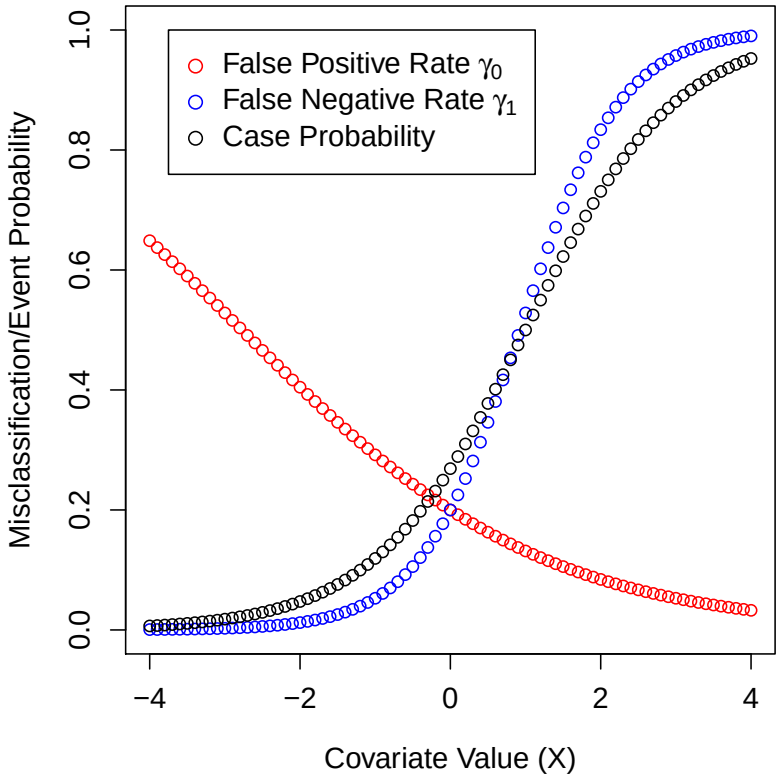

Figure 7. Graphical display of false positive, false negative and event probabilities for differential misclassification scenario 1 in which the false positive rate increases with the probability of being a control and false negative rate increases with the probability of being a case. The misclassification in this scenario leads to an underestimate of the true AUC value.

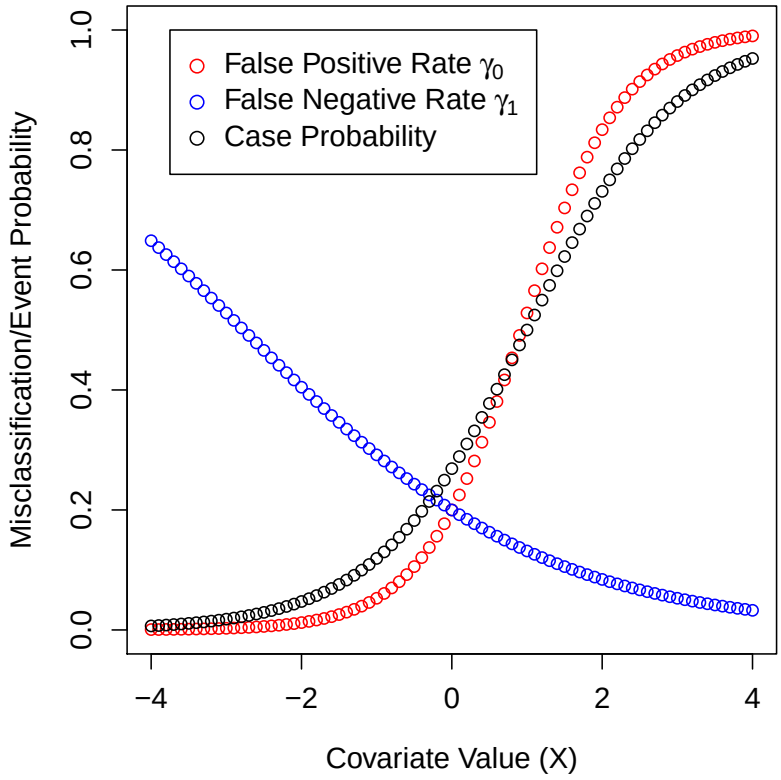

Figure 8. Graphical display of false positive, false negative and event probabilities for differential misclassification scenario 2 in which the false positive rate increases with the probability of being a case and the false negative rate increases with the probability of being a control. The misclassification in this scenario leads to an overestimate of the true AUC value.

Statistics
in Medicine

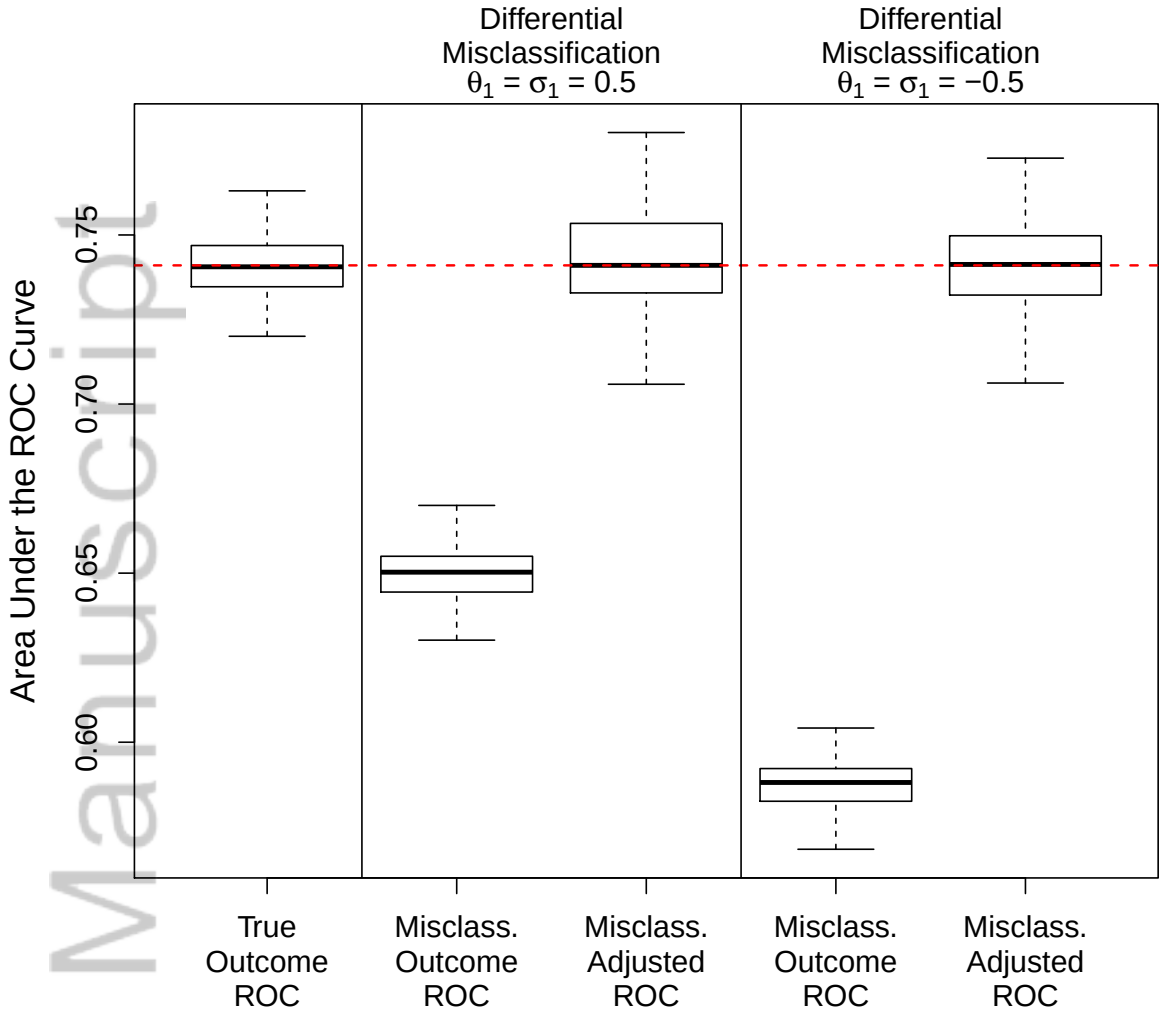

Figure 9. Examples of differential misclassification in which false positive and false negative rates have same direction. When $\theta_1 = \sigma_1 = 0.5$, both the false positive and false negative rates increase with the probability of being a case and the misclassification leads to an underestimate of the true AUC. When $\theta_1 = \sigma_1 = -0.5$, both the false positive and false negative rates decrease with the probability of being a case and, again, the misclassification leads to an underestimate of the true AUC. In both cases, the misclassification-adjusted ROC method substantially reduces AUC bias.

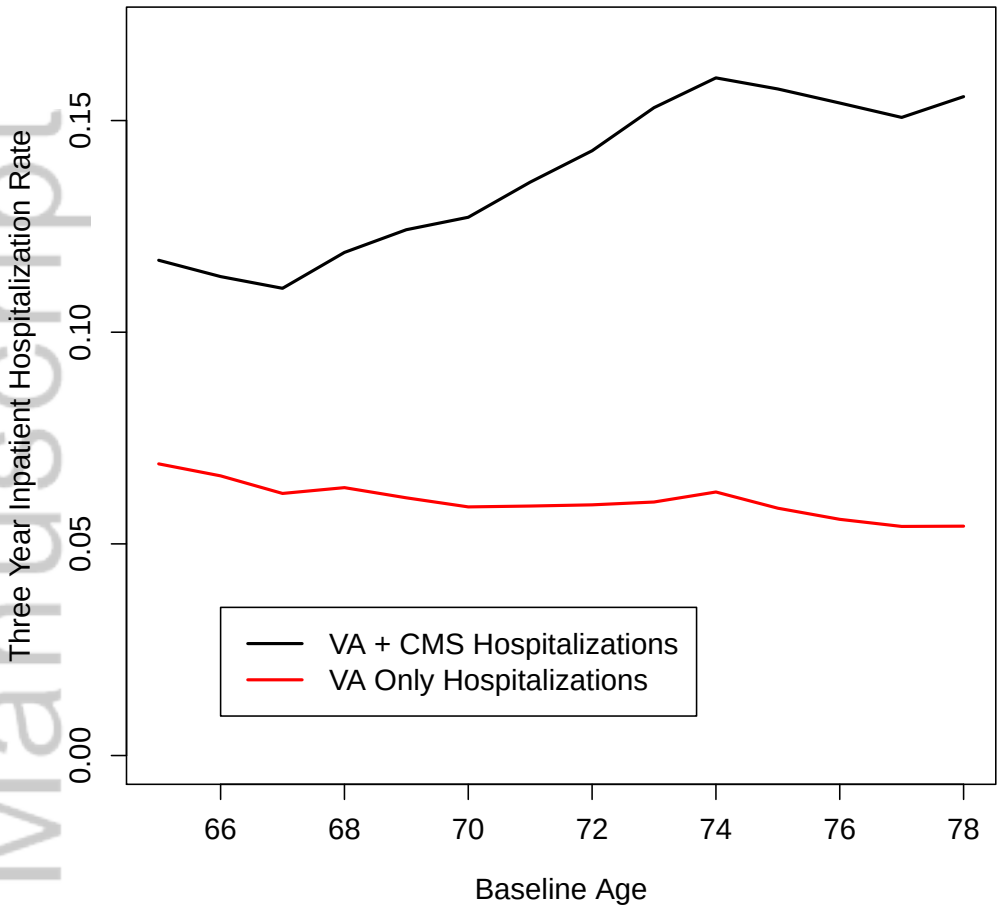

Figure 10. Three year inpatient hospitalization rates by baseline age using events recorded in either VHA or CMS records (black line) and only VHA EHR events (red line). Using only the VHA EHR leads to a misclassification of hospitalization outcome events as non-events. The change in difference between event rates by age indicates covariate dependent misclassification.

Statistics
in Medicine

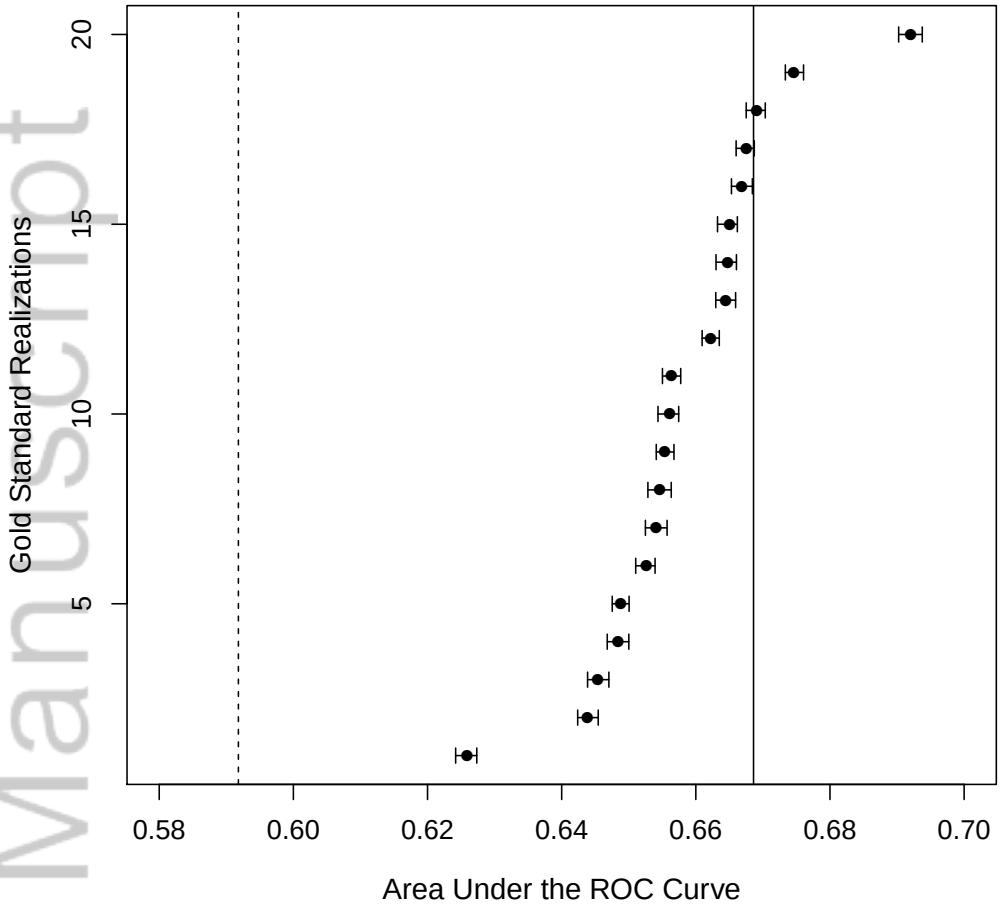

Figure 11. Bootstrap-based 90% confidence intervals of AUC_M for 20 realizations of the internal validation cohort in the data example. The solid black line at AUC=0.669 indicates the true AUC value and the dashed line at AUC=0.592 the nave analysis of misclassified outcomes. Coverage of the bootstrap-based confidence intervals suffered due to the misclassification rates being estimated rather than known exactly.

Baseline Risk $e^{\beta_0}/(1+e^{\beta_0})$	Effect Size β_1	False Positive Rate γ_0	False Negative Rate γ_1	True Outcome ROC $AUC(T,\hat{P})$	Misclass. Outcome ROC $AUC(Y,\hat{P^I})$	Mean Bias (% Change)
0.01	0.1	0.05	0.05	0.511	0.499	1.60
			0.2	0.511	0.500	1.31
		0.2	0.05	0.511	0.499	1.56
			0.2	0.511	0.500	1.27
	1.0	0.05	0.05	0.758	0.560	26.07
			0.2	0.758	0.552	27.16
		0.2	0.05	0.758	0.517	31.77
			0.2	0.758	0.512	32.44
0.2	0.1	0.05	0.05	0.527	0.522	0.95
			0.2	0.527	0.520	1.28
		0.2	0.05	0.527	0.513	2.61
			0.2	0.527	0.510	3.10
	1.0	0.05	0.05	0.743	0.704	5.21
			0.2	0.743	0.688	7.42
		0.2	0.05	0.743	0.641	13.72
			0.2	0.743	0.617	16.87

Table 1. Bias in Area Under the ROC curve (AUC) for misclassified outcomes in a population cohort in which controls outnumber cases. The table shows the AUC for standard ROC analysis of true and misclassified outcomes simulated at low and high values for each model parameter. The bias in AUC is greatest for very rare events with large effect covariates. In this scenario, misclassifying controls as cases (false positives) has a greater impact on AUC bias than does misclassifying true cases. Supplementary Table 2 shows the result when cases are more prevalent in the cohort.

Baseline Risk $e^{\beta_0}/(1+e^{\beta_0})$	Effect Size β_1	False Positive Rate γ_0	False Negative Rate γ_1	True Outcome ROC $AUC(T,\hat{P})$	Misclass. Outcome ROC $AUC(Y,\hat{P^I})$	Mean Bias (% Change)
0.6	0.1	0.05	0.05	0.527	0.524	0.57
			0.2	0.527	0.519	1.53
		0.2	0.05	0.527	0.521	1.13
			0.2	0.527	0.515	2.29
	1.0	0.05	0.05	0.739	0.714	3.41
			0.2	0.739	0.675	8.75
		0.2	0.05	0.739	0.689	6.87
			0.2	0.739	0.641	13.30
0.8	0.1	0.05	0.05	0.528	0.523	0.99
			0.2	0.528	0.514	2.70
		0.2	0.05	0.528	0.520	1.47
			0.2	0.528	0.510	3.42
	1.0	0.05	0.05	0.742	0.704	5.15
			0.2	0.742	0.640	13.76
		0.2	0.05	0.742	0.687	7.39
			0.2	0.742	0.618	16.79

Table 2. Bias in Area Under the ROC curve (AUC) for misclassified outcomes in a population cohort in which cases outnumber controls. In this scenario, misclassifying cases as controls (false negative) has a greater impact on AUC bias than does misclassifying true controls.