CORRECTED ROC ANALYSIS FOR MISCLASSIFIED BINARY OUTCOMES[1]

CHEN CHEN, RACHEL GIBLON

DECEMBER 6, 2022

1. Zawistowski M, Sussman JB, Hofer TP, Bentley D, Hayward RA, Wiitala WL. Corrected ROC analysis for misclassified binary outcomes. Statistics in Medicine. 2017 Jun 15;36(13):2148-60.

Introduction & Motivation

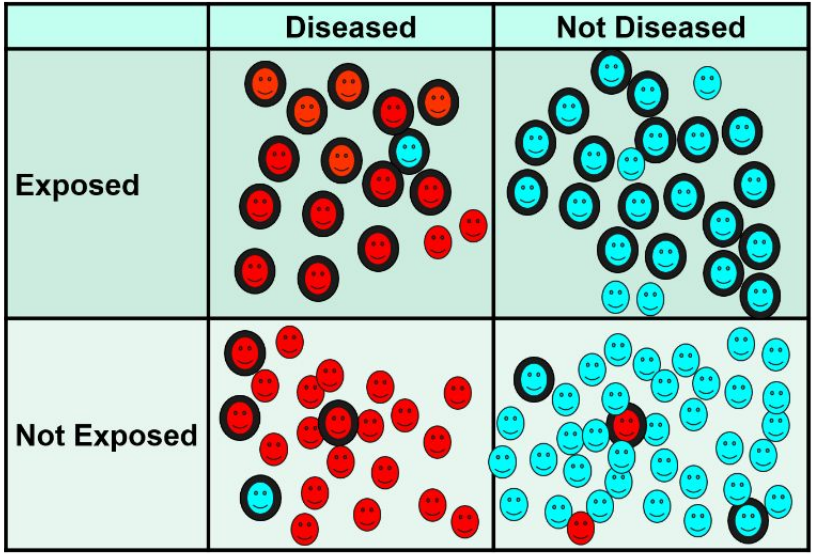
Methods

Results

Conclusions

INTRODUCTION & MOTIVATION

- Misclassification outcome is a common systematic error in EHR
  - Incorrect diagnosis, screening test, etc
- Ignoring misclassification in outcome:
  - Regression estimators will be biased
  - Loss of efficiency
  - violates the assumption for ROC analysis: binary outcomes are classified correctly



Aim:Correct ROC analysis to account for misclassification bias and increase accuracy and obtained bias-corrected AUC

LITERATURE REVIEW

Authors and Year	Correction method	Advantage	Disadvantage
Neuhaus (1999)	Modified likelihood function to account misclassification rate	Consistent estimators to reduce bias	Less efficient, assume misclassification rate is known
Magder and Hughes (1997)	EM algorithms	Applicable to case-control study and other forms of binary regression; accommodate differential misclassification;	Identifiability issue when specificity and sensitivity are unknown
Mcinturff et al (2004)	Bayesian method: Beta priors to sensitivity and specificity. Different magnitude compared to Magder and Hughes.	Include prior information for covariates	Computation heavy

METHODS- modified likelihood

- Misclassification rate: false positive and false negative

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

T represent true unobserved outcome and Y represent observed misclassified outcome

- Standard logistics regression model

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

- Misclassification-adjusted likelihood

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

METHODS- misclassification-adjusted predictive probability

- Predictive probability using misclassification-adjusted estimator

$\gamma_1(\mathbf{X}) = P(Y = 0|T = 1, \mathbf{X}).$

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

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

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

METHODS- misclassification adjusted ROC

Standard ROC

$$ROC(\alpha, \mathbf{B}, \mathbf{q}) = \left(FP(\alpha, \mathbf{B}, \mathbf{q}), TP(\alpha, \mathbf{B}, \mathbf{q})\right)$$

$$AUC(\mathbf{B}, \mathbf{q}) = \int_{\alpha} ROC(t, \mathbf{B}, \mathbf{q})dt$$

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

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

B is a sets of general outcome and q is risk prediction score

Misclassification adjusted ROC

$$ROC_M(\alpha) = (FP_M(\alpha), TP_M(\alpha))$$

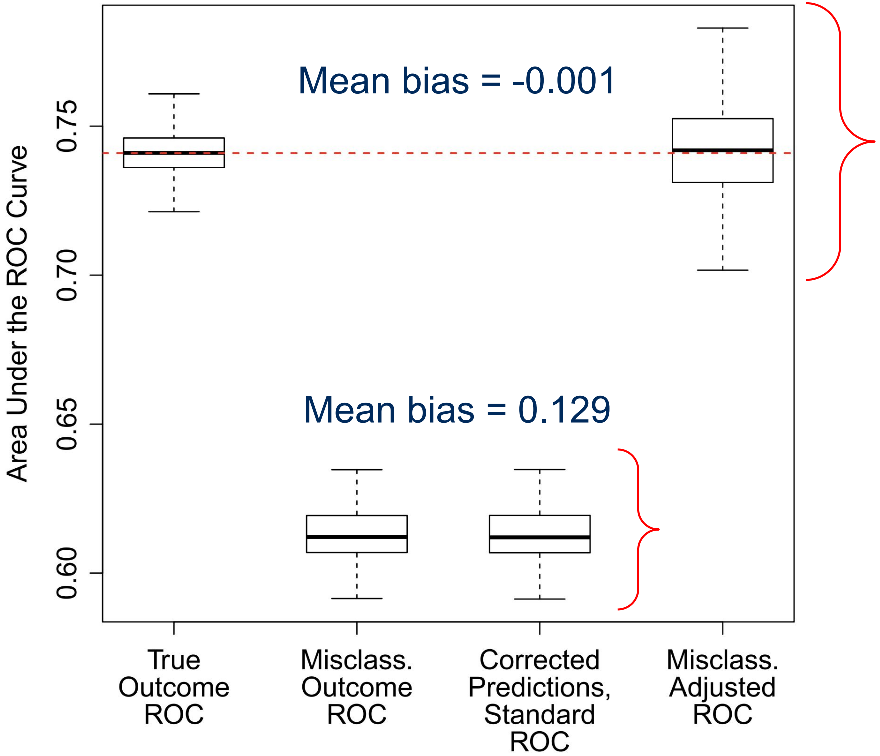
$$AUC_M = \int_{\alpha} ROC_M(t)dt,$$

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

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

RESULTS: SIMULATIONS

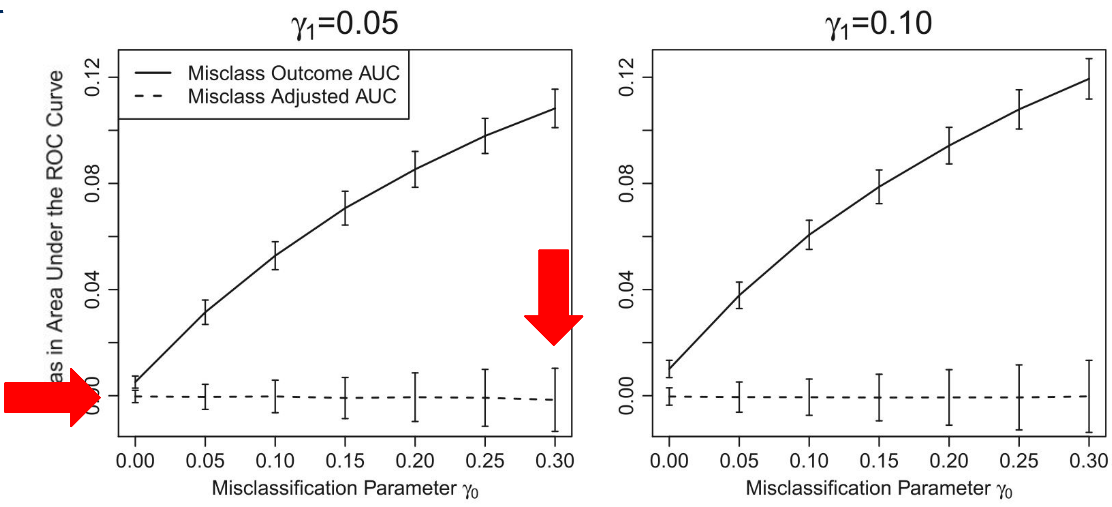
- ROC adjustment removed almost all bias in the AUC
- Larger variance in bootstrap-based confidence intervals



UNIVERSITY OF TORONTO

RESULTS: SIMULATIONS

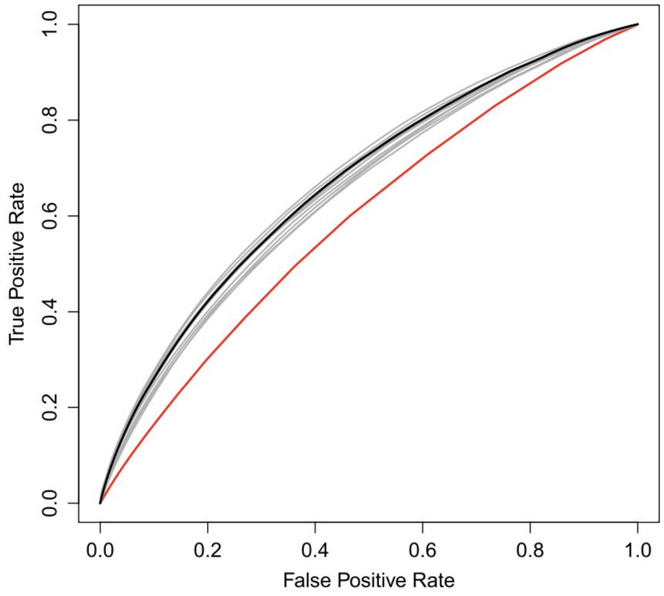
- ROC adjustment removed almost all bias in the AUC
- Larger variance in bootstrap-based confidence intervals
- Bias remained close to zero over a range of FP and FN misclassification rates
- SE on adjusted AUC estimates was sensitive to increases in misclassification



UNIVERSITY OF TORONTO

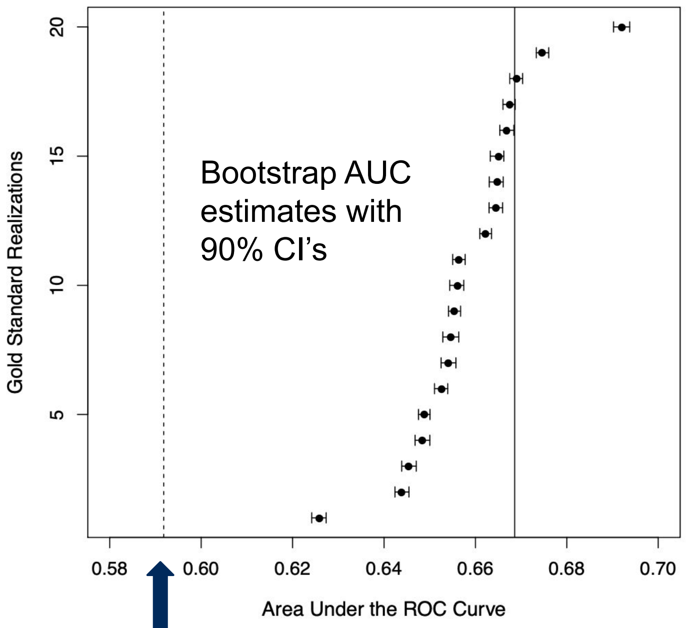
RESULTS: REAL DATA

ROC Curves



90% CI Coverage

True AUC value



Naive estimate

UNIVERSITY OF TORONTO

CONCLUSIONS

- Misclassified (binary) outcomes → biased AUC estimates
- Existing methods focus on parameter estimates of the regression model
- Corrected ROC analysis addresses this problem
- Computationally straightforward (sample R & STATA code provided)

- Assumptions of this method:
  - Misclassification rates are known or can be inferred
  - Sample size is large enough for maximum likelihood estimates