The Analysis of Placement Values for Evaluating Discriminatory Measures

Margaret Sullivan Pepe & Tianxi Cai

Biometrics (2004)

Allison Meisner · May 27, 2014

Overview

When we have a continuous test Y and a binary outcome D, the ROC curve plots the (FPR, TPR) pairs for each possible cutoff of the test.

Problem: The ROC curve may differ by patient characteristics. Identifying such variability helps us to apply the test in an optimal way.

Solution: ROC regression **with placement values**

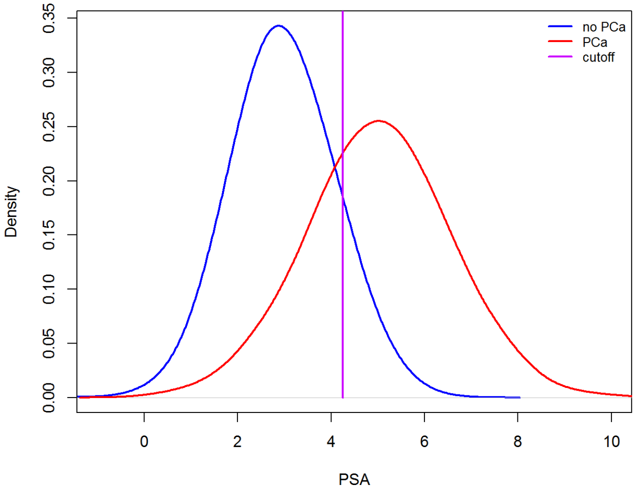
Motivating Example

Prostate-specific antigen (PSA) is a popular, though controversial, way to screen men for prostate cancer (PCa).

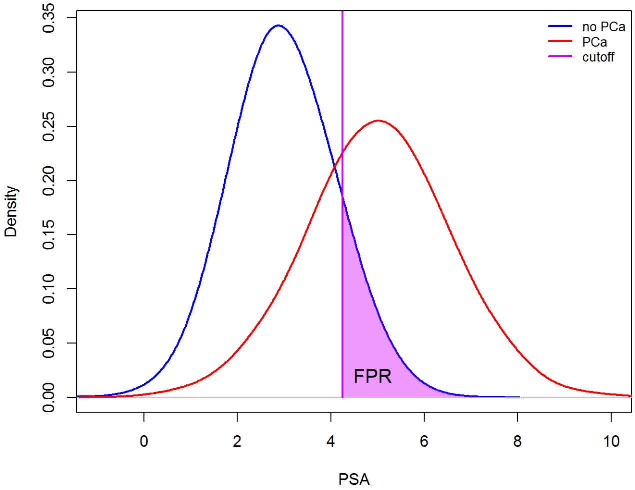
The biology of PSA and PCa has implications for the usefulness of PSA as a screening tool:

- PSA levels differ by age: older men typically have higher PSA, regardless of PCa status
- Age can potentially affect the ability of PSA to discriminate PCa cases
- Among PCa cases, PSA measured closer to diagnosis does a better job of discriminating PCa

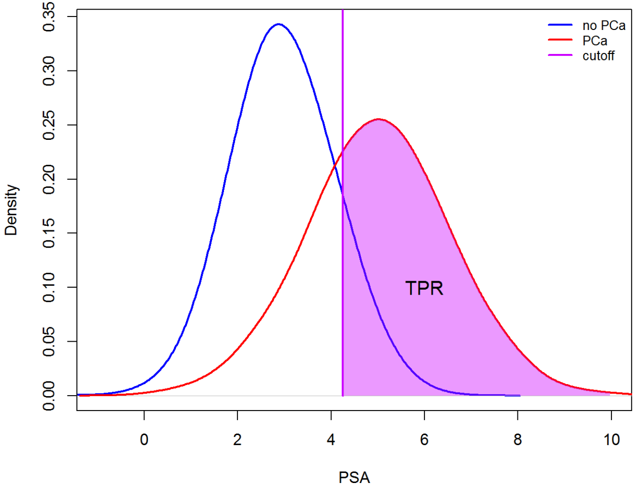
Background: FPR, TPR, ROC



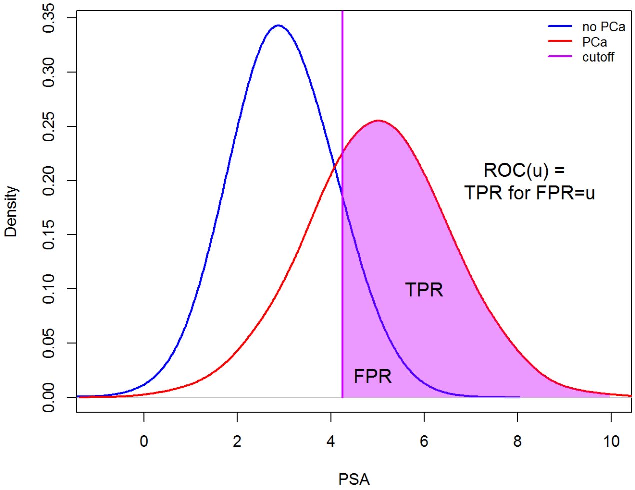
Background: FPR, TPR, ROC



Background: FPR, TPR, ROC

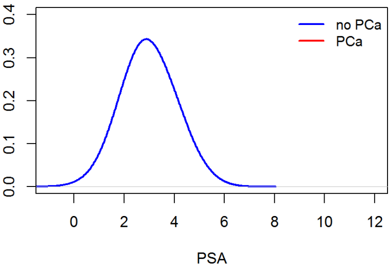


Background: FPR, TPR, ROC

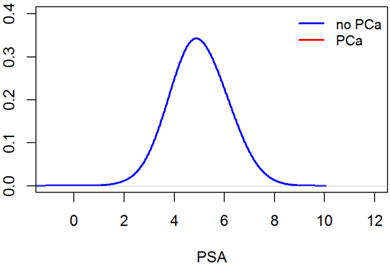


Background: Effect of Covariates on ROC

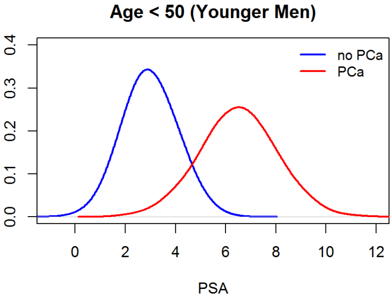
Age < 50 (Younger Men)

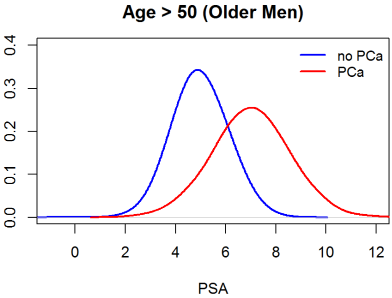


Age > 50 (Older Men)

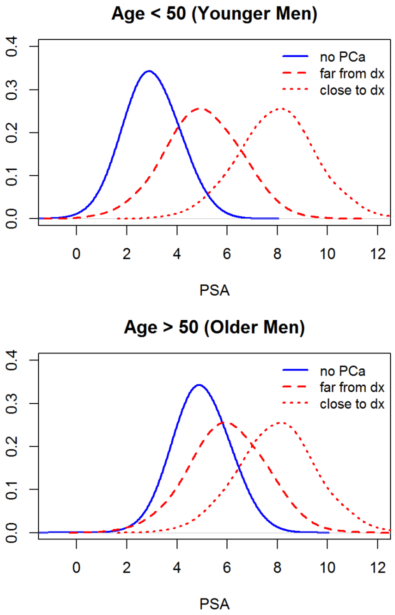


Background: Effect of Covariates on ROC

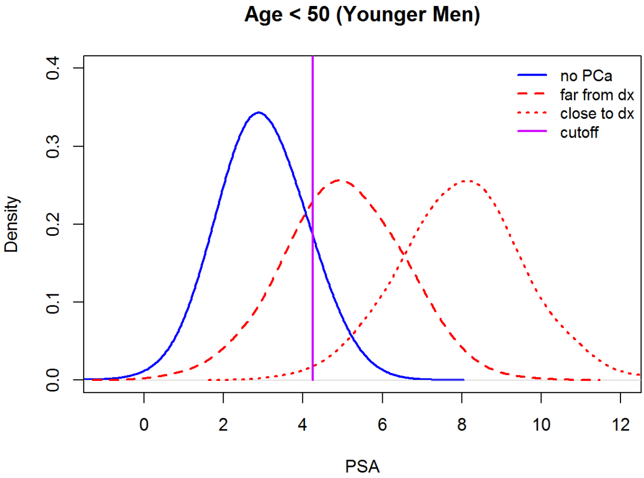




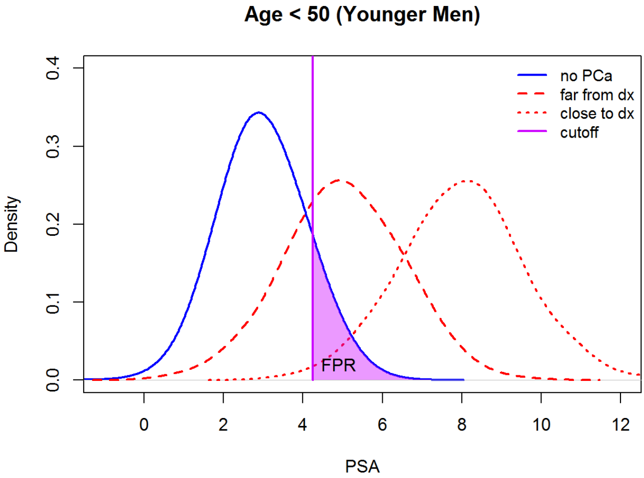
Background: Effect of Covariates on ROC



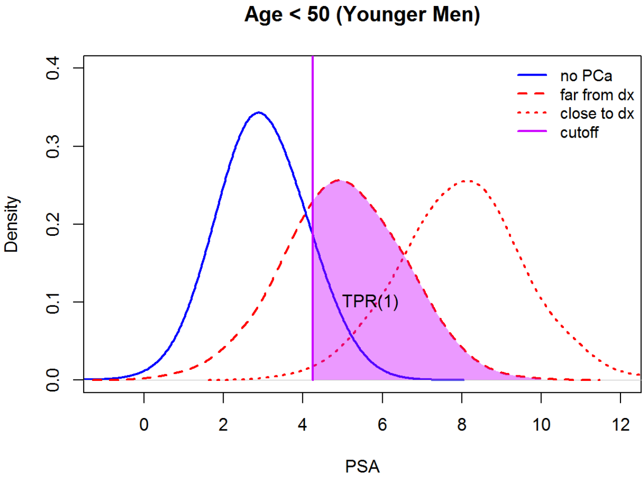
Background: Effect of Covariates on ROC



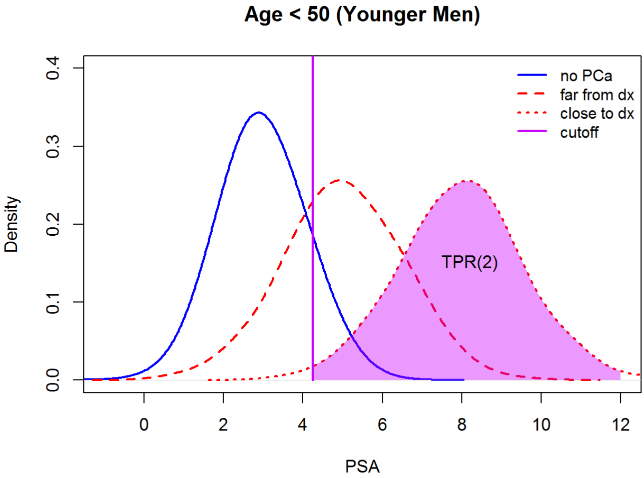
Background: Effect of Covariates on ROC



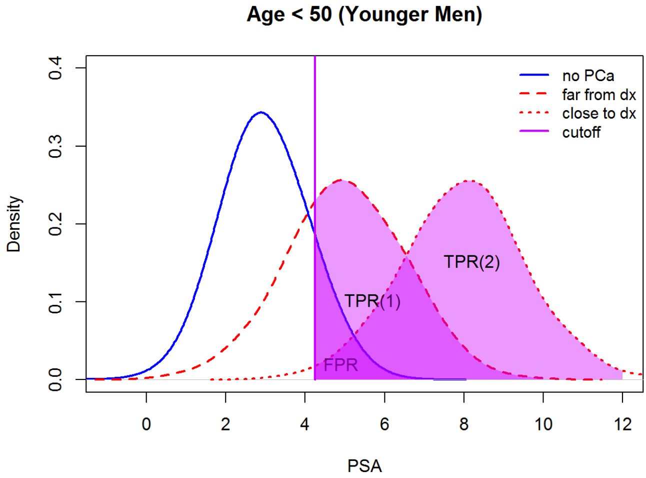
Background: Effect of Covariates on ROC



Background: Effect of Covariates on ROC



Background: Effect of Covariates on ROC



Recall, $ROC(u) = (\text{TPR at FPR} = u)$.

ROC Model

- ROC model (Pepe, 1997): $ROC_{\mathbf{Z}_D}(u) = g(\boldsymbol{\beta}^T \mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u))$
  - $\boldsymbol{\alpha}$ = underlying shape of ROC curve
  - $\boldsymbol{\beta}$ = impact of $\mathbf{Z}_D$ on shape of ROC curve
- Problem: estimation
  - Pepe (2000) and Alonzo and Pepe (2002) create indicators $I(Y_{Di} \geq F_{\overline{D}}^{-1}(1-u))$ for some set of FPRs u and then use binary regression techniques
  - Pepe & Cai propose using placement values and what is known about their distribution to estimate the parameters more efficiently

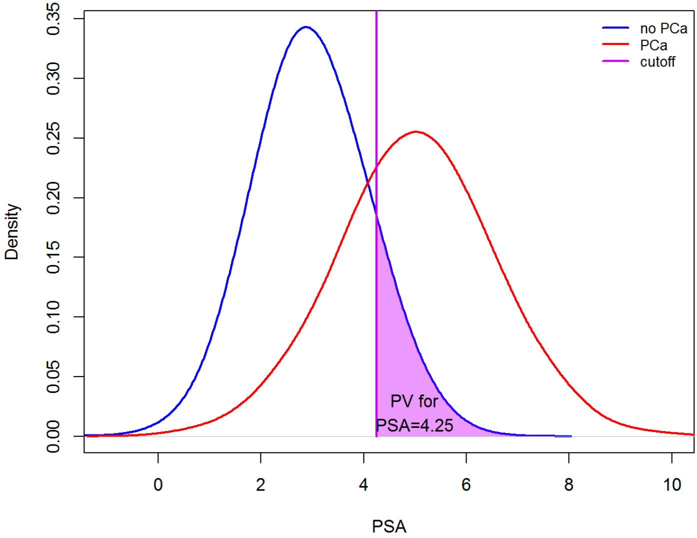
Placement Values

- Definitions
  - Placement values: $U_{Di} = 1 - F_{\overline{D}}(Y_{Di})$ for the i^{th} diseased subject. In words, the placement value for the i^{th} diseased subject is the proportion of the reference (non-diseased) population with marker Y values above Y_{Di}.
    - If $\mathbf{Z}_{\overline{D}}$ affects the distribution of Y in the reference population, $U_{Di} = 1 - F_{\overline{D},\mathbf{Z}_{\overline{D}}}(Y_{Di})$.
  - ROC curve: $ROC(u) = P(Y_D \geq F_{\overline{D}}^{-1}(1-u))$ = (TPR at FPR=u)

- Relationship between ROC and placement values

$$
\begin{aligned}
ROC(u) &= P(Y_D \geq F_{\overline{D}}^{-1}(1-u)) = P(1-u \leq F_{\overline{D}}(Y_D)) \\
&= P(1 - F_{\overline{D}}(Y_D) \leq u) = P(U_D \leq u)
\end{aligned}
$$

Placement Values



Proposed Method

- ROC model (Pepe, 1997): $ROC_{\mathbf{Z}_D}(u) = g(\boldsymbol{\beta}^T \mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u))$
- Proposed model: $H_{\boldsymbol{\alpha}}(U_D) = -\boldsymbol{\beta}^T \mathbf{Z}_D + \epsilon$, where $\epsilon \sim g$
- Proof of equivalence:

$$
\begin{aligned}
Pr(U_D \leq u) &= Pr(H_{\boldsymbol{\alpha}}(U_D) \leq H_{\boldsymbol{\alpha}}(u)) \\
&= Pr(-\boldsymbol{\beta}^T \mathbf{Z}_D + \epsilon \leq H_{\boldsymbol{\alpha}}(u)) \\
&= Pr(\epsilon \leq \boldsymbol{\beta}^T \mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u)) \\
&= g(\boldsymbol{\beta}^T \mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u)) = ROC_{\mathbf{Z}_D}(u)
\end{aligned}
$$

  Recall that if $\mathbf{Z}_{\overline{D}}$ affects the distribution of Y in the reference population, $U_{Di} = 1 - F_{\overline{D},\mathbf{Z}_{\overline{D}}}(Y_{Di})$; then we may write

$$H_{\boldsymbol{\alpha}}(U_D) = -\boldsymbol{\beta}^T \mathbf{Z}_D + \epsilon \; \Leftrightarrow \; ROC_{\mathbf{Z}_{\overline{D}},\mathbf{Z}_D}(u) = g(\boldsymbol{\beta}^T \mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u))$$

- In our example, $\mathbf{Z}_{\overline{D}}$ = age and $\mathbf{Z}_D$ = (age, time).

Proposed Method: Algorithm

Since $Pr(U_D \leq u) = g(\boldsymbol{\beta}^T \mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u))$, we know the density function is

$$f(u) = \frac{\partial g(\boldsymbol{\beta}^T \mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u))}{\partial u}.$$

Then, for $[a, b] \subset (0, 1)$, the log likelihood is

$$\begin{aligned} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n_D} [& I(U_{Di} < a) \log\{g(\boldsymbol{\beta}^T \mathbf{Z}_{Di} + H_{\boldsymbol{\alpha}}(a))\} \\ & + I(U_{Di} > b) \log\{1 - g(\boldsymbol{\beta}^T \mathbf{Z}_{Di} + H_{\boldsymbol{\alpha}}(b))\} \\ & + I(U_{Di} \in (a, b)) \log f(U_{Di})] \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Proposed Method: Algorithm

Estimating $F_{\overline{D},\mathbf{Z}_{\overline{D}}}$

- Pepe and Cai advise estimating $F_{\overline{D},\mathbf{Z}_{\overline{D}}}$ nonparametrically if $\mathbf{Z}_{\overline{D}}$ is discrete and semiparametrically otherwise.
- For semiparametric estimation, Pepe and Cai recommend the semiparamtric regression quantile estimation procedure developed by Heagerty and Pepe (1999).

The estimates of the placement values, $\hat{U}_{Di}$, are substituted into $\ell(\boldsymbol{\theta})$, yielding a pseudo-log-likelihood*, which is maximized to estimate $\boldsymbol{\theta}$.

Competing Method: Algorithm

Alonzo and Pepe proposed an algorithm for fitting ROC regression based on binary regression methods.

1. For $[a, b] \subset (0, 1)$, let

$$T = \{u_1, ..., u_{n_T}\} = \{1 - j/n_{\overline{D}};\ \ j = 1, ..., n_{\overline{D}} - 1\} \cap [a, b]$$

   (the maximal set).

2. Then for each diseased subject i, the n_T binary variables B_{ui} are calculated:

$$B_{ui} = I[\hat{U}_{Di} \leq u],\ \ u \in T.$$

3. The binary generalized linear regression model

$$E\{B_{ui}\} = g\{\boldsymbol{\beta}^T \mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u)\}$$

   is fit using standard techniques.

The Pepe and Cai method is claimed to be **more efficient** than that of Alonzo and Pepe.

Simulations

Set-up

- $Y_D = \alpha_1^{-1}\{\alpha_0 + \beta_1 Z_1 + (\beta_2 + 0.5\alpha_1)Z_2 + \epsilon_D\}$
  $Y_{\overline{D}} = 0.5Z_2 + \epsilon_{\overline{D}}$
- $Z_1 \sim$ Bernoulli(0.5), $Z_2 \sim$ Uniform$(0,1)$
- $\epsilon_D \sim N(0,1)$, $\epsilon_{\overline{D}} \sim N(0,1)$

Induced ROC curve:

$$\begin{aligned}
ROC_{\mathbf{Z}_{\overline{D}},\mathbf{Z}_D}(u) &= Pr(U_D \leq u) = Pr(1 - F_{\overline{D}}(Y_D) \leq u) \\
&= Pr(F_{\overline{D}}^{-1}(1-u) \leq \alpha_1^{-1}\{\alpha_0 + \beta_1 z_1 + (\beta_2 + 0.5\alpha_1)z_2 + \epsilon_D) \\
&= Pr(\Phi^{-1}(1-u) + 0.5z_2 \leq \\
&\quad \alpha_1^{-1}\{\alpha_0 + \beta_1 z_1 + (\beta_2 + 0.5\alpha_1)z_2 + \epsilon_D\}) \\
&= Pr(\epsilon_D \leq -\alpha_1\Phi^{-1}(1-u) + \alpha_0 + \beta_1 z_1 + \beta_2 z_2) \\
&= \Phi(\alpha_1\Phi^{-1}(u) + \alpha_0 + \beta_1 z_1 + \beta_2 z_2) = g(\beta^T\mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u))
\end{aligned}$$

Recall, $\boldsymbol{\alpha}$ = shape of ROC, $\boldsymbol{\beta}$ = effects of $\mathbf{Z}_D$ on ROC

Simulations

Note that here

$$\mathbf{Z}_{\overline{D}} = Z_2 \text{ and } \mathbf{Z}_D = (Z_1, Z_2).$$

Despite their recommendations, Pepe and Cai did not use the semiparametric method of Heagerty and Pepe to estimate placement values.

Instead, Pepe and Cai regress Y on Z_2 among the non-diseased subjects:

$$E(Y_{\overline{D}}|Z_2 = z_2) = \gamma_0 + \gamma_1 z_2 \;\Rightarrow\; \hat{\epsilon}_{\overline{D}i} = Y_{\overline{D}i} - \hat{\gamma}_0 - \hat{\gamma}_1 z_{2\overline{D}i}.$$

Then the placement value for subject i was estimated to be

$$\hat{U}_{Di} = \frac{1}{n_{\overline{D}}} \sum_{j=1}^{n_{\overline{D}}} I(\hat{\epsilon}_{\overline{D}_j} > Y_{Di} - \hat{\gamma}_0 - \hat{\gamma}_1 z_{2Di}).$$

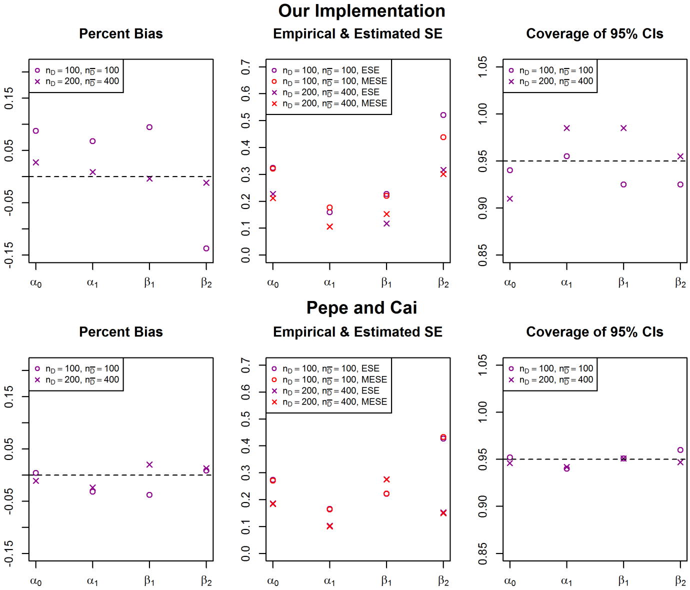
Simulations

Two sets of simulations (1000 simulations each):

1. Pepe and Cai method only
   - Bias
   - Empirical SE
   - Mean estimated SE
   - Empirical coverage probability
   - Note: $\alpha_0 = 1, \alpha_1 = 1, \beta_1 = 0.5, \beta_2 = 0.7$ throughout
   - Considered $[a, b] = [0.01, 0.99]$ and $[a, b] = [0.01, 0.20]$
2. Pepe and Cai vs. Alonzo and Pepe
   - Bias
   - MSE
   - Two sets of parameter values considered
     - $\alpha_0 = 1, \alpha_1 = 1, \beta_1 = 0.5, \beta_2 = 0.7$
     - $\alpha_0 = 1.5, \alpha_1 = 0.9, \beta_1 = 0.5, \beta_2 = 0.7$
   - Considered $[a, b] = [0.01, 0.99]$ and $[a, b] = [0.01, 0.50]$

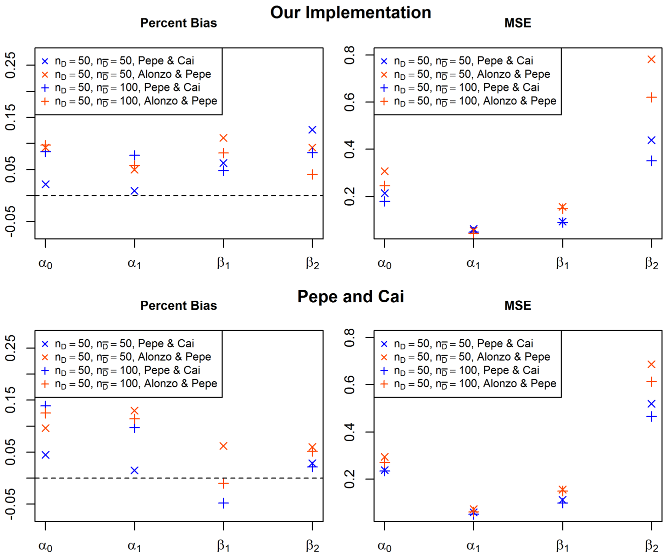
Simulations: Pepe & Cai

- $[a, b] = [0.01, 0.99]$



Simulations: Pepe & Cai vs. Alonzo & Pepe

- $\alpha_0 = 1, \alpha_1 = 1, \beta_1 = 0.5, \beta_2 = 0.7$
- $[a, b] = [0.01, 0.99]$

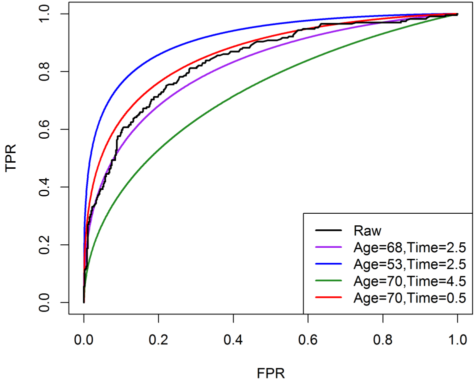


Application

The proposed method was applied to data from a study on PSA and PCa screening.

- 88 PCa cases, 88 age-matched controls
- Recall, $\mathbf{Z}_{\overline{D}}$ = age and $\mathbf{Z}_D$ = (age, time)
- Model: $ROC_{\mathbf{Z}_{\overline{D}},\mathbf{Z}_D}(u) = \Phi(\alpha_0 + \alpha_1\Phi^{-1}(u) + \beta_1\text{time} + \beta_2\text{age})$
- SE estimates from the bootstrap (500 replications)

	Estimate (SE)
α_0	4.30 (0.93)
α_1	0.84 (0.09)
β_1	-0.16 (0.03)
β_2	-0.04 (0.01)



Conclusions

- The proposed method has nice intuition behind it and makes full use of the data through placement values, as opposed to creating indicators.
- Implementation of the proposed method is less straightforward and is not particularly computationally efficient.
- In most scenarios, the proposed method is more statistically efficient than the binary regression technique.
- Both methods are susceptible to misspecification in both the estimation of $F_{\overline{D}}$ and the form of the ROC model.

Effects of Misspecification

What happens when

$$Y_{\overline{D}} = 0.5Z_2^2 + N(0, (Z_2 + 0.5)^2)$$

but we still assume

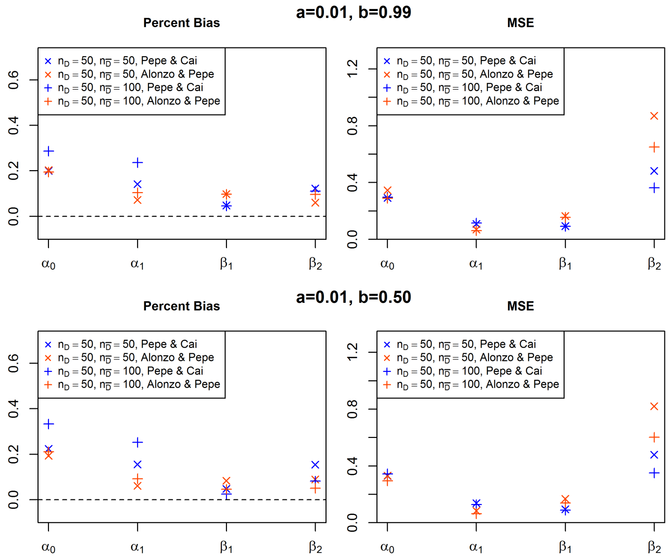
$$Y_{\overline{D}} = 0.5Z_2 + N(0, 1)?$$

This will impact

1. estimates of placement values
2. form of the induced ROC curve (used in the likelihood calculation)

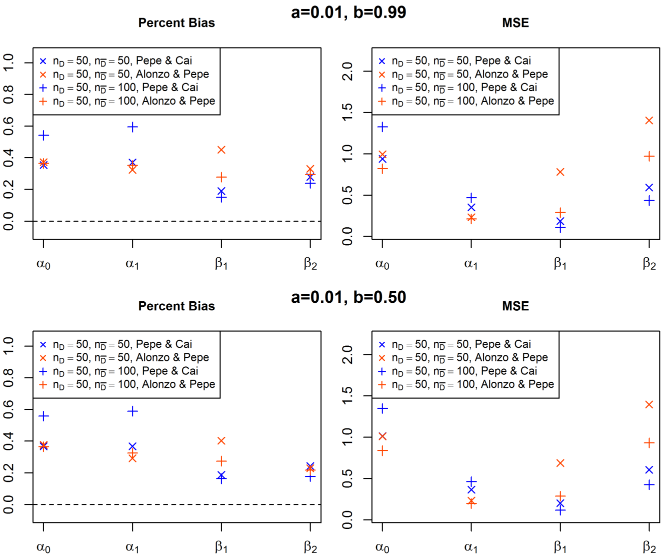
Effects of Misspecification

- $\alpha_0 = 1, \alpha_1 = 1, \beta_1 = 0.5, \beta_2 = 0.7$



Effects of Misspecification

- $\alpha_0 = 1.5, \alpha_1 = 0.9, \beta_1 = 0.5, \beta_2 = 0.7$

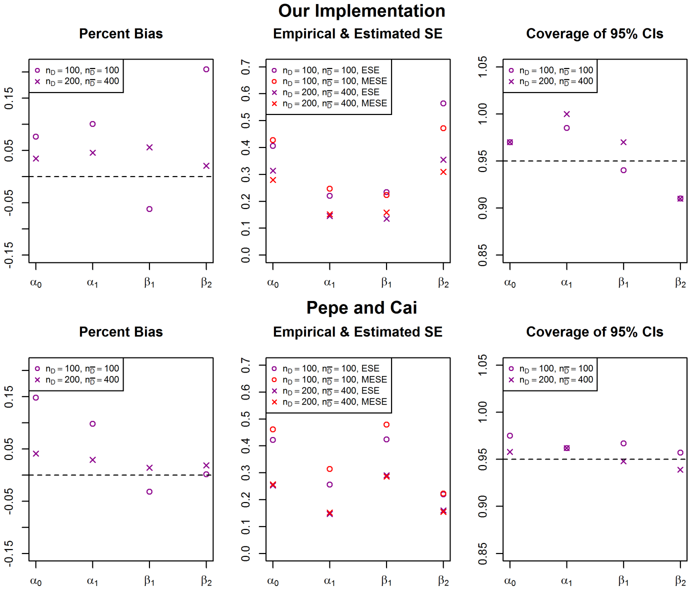


Conclusions

- The proposed method has nice intuition behind it and makes full use of the data through placement values, as opposed to creating indicators.
- Implementation of the proposed method is less straightforward and is not particularly computationally efficient.
- In most scenarios, the proposed method is more statistically efficient than the binary regression technique.
- Both methods are susceptible to misspecification in both the estimation of $F_{\overline{D}}$ and the form of the ROC model.

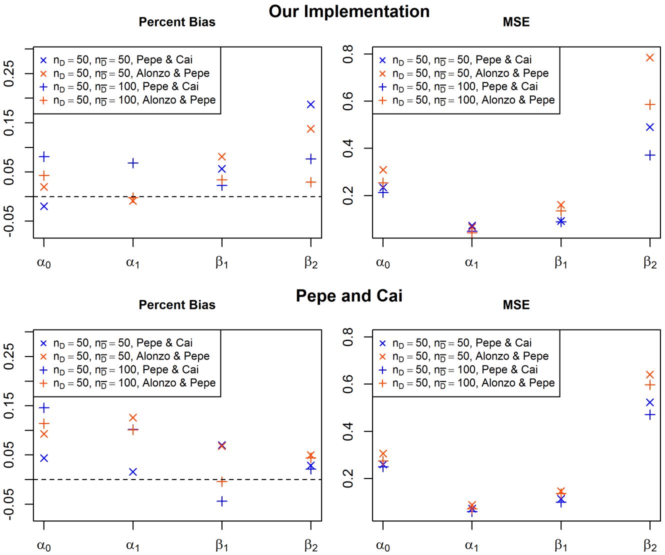
Simulations: Pepe & Cai

- $[a, b] = [0.01, 0.20]$



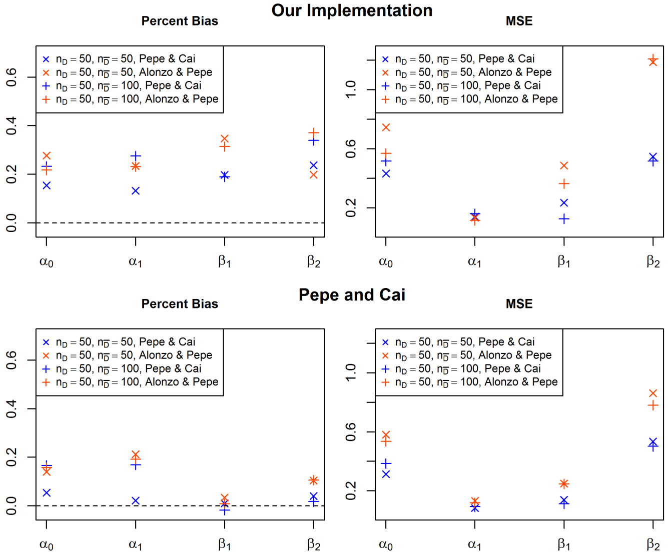
Simulations: Pepe & Cai vs. Alonzo & Pepe

- $\alpha_0 = 1, \alpha_1 = 1, \beta_1 = 0.5, \beta_2 = 0.7$
- $[a, b] = [0.01, 0.50]$



Simulations: Pepe & Cai vs. Alonzo & Pepe

- $\alpha_0 = 1.5, \alpha_1 = 0.9, \beta_1 = 0.5, \beta_2 = 0.7$
- $[a, b] = [0.01, 0.99]$



Simulations: Pepe & Cai vs. Alonzo & Pepe

- $\alpha_0 = 1.5, \alpha_1 = 0.9, \beta_1 = 0.5, \beta_2 = 0.7$
- $[a, b] = [0.01, 0.0.5]$