The Analysis of Placement Values for Evaluating Discriminatory Measures

Margaret Sullivan Pepe & Tianxi Cai
Biometrics (2004)

Allison Meisner · May 6, 2014

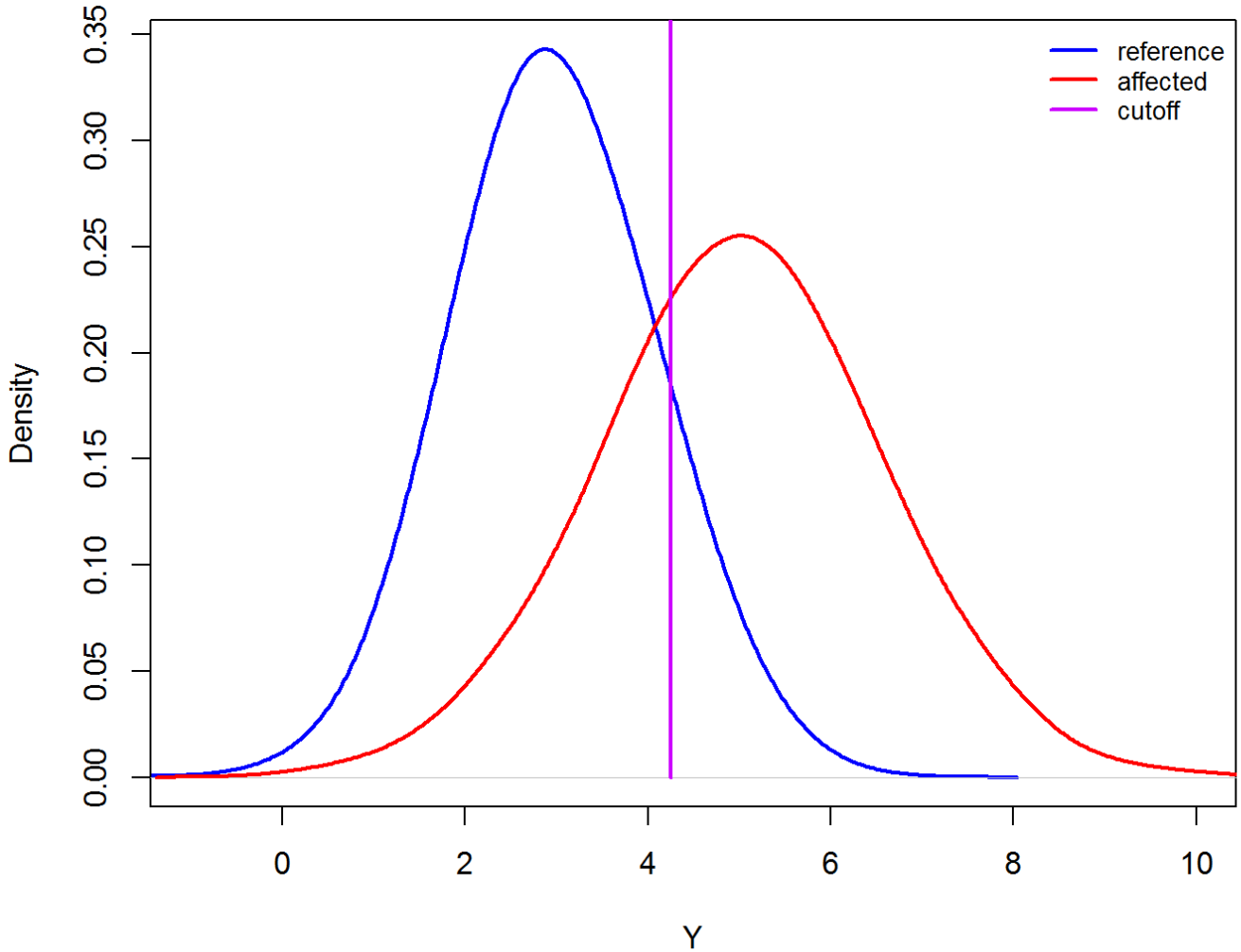
Quick Review

Recall that 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.

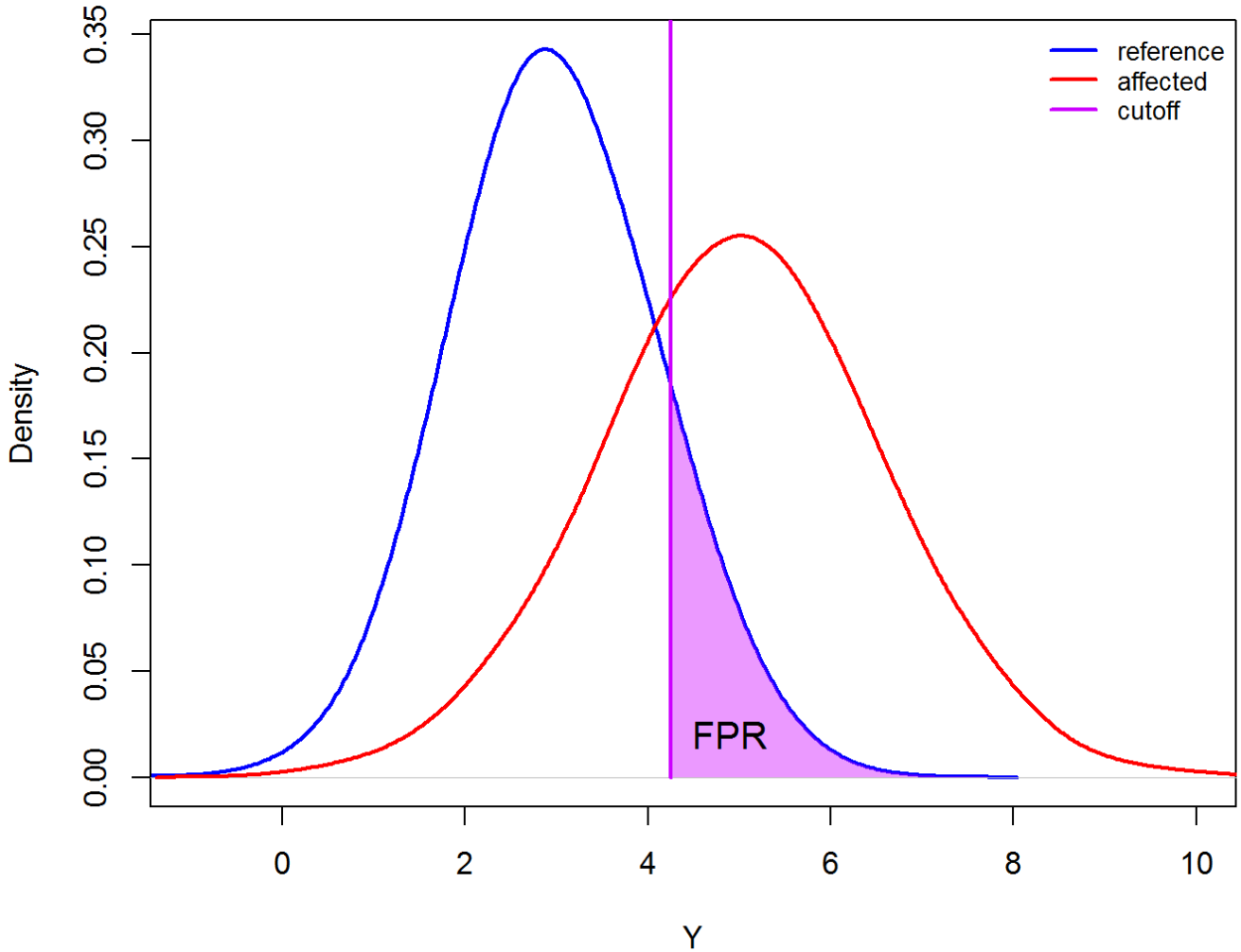
Problem: The ROC curve may differ by patient characteristics. Identifying such variability helps us to apply the test in an optimal way.

Solution: ROC regression **with placement values (PVs)**

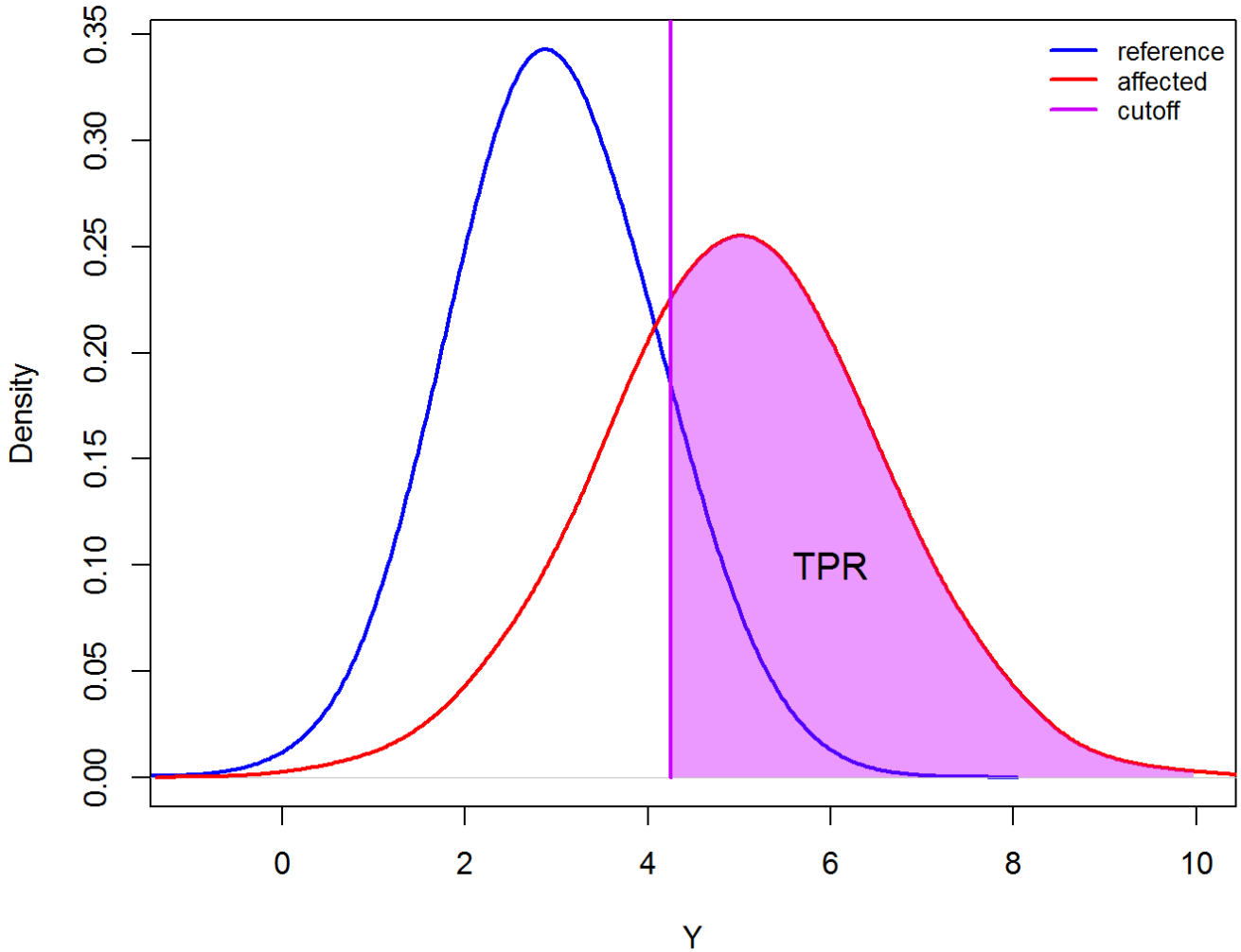
Quick Review



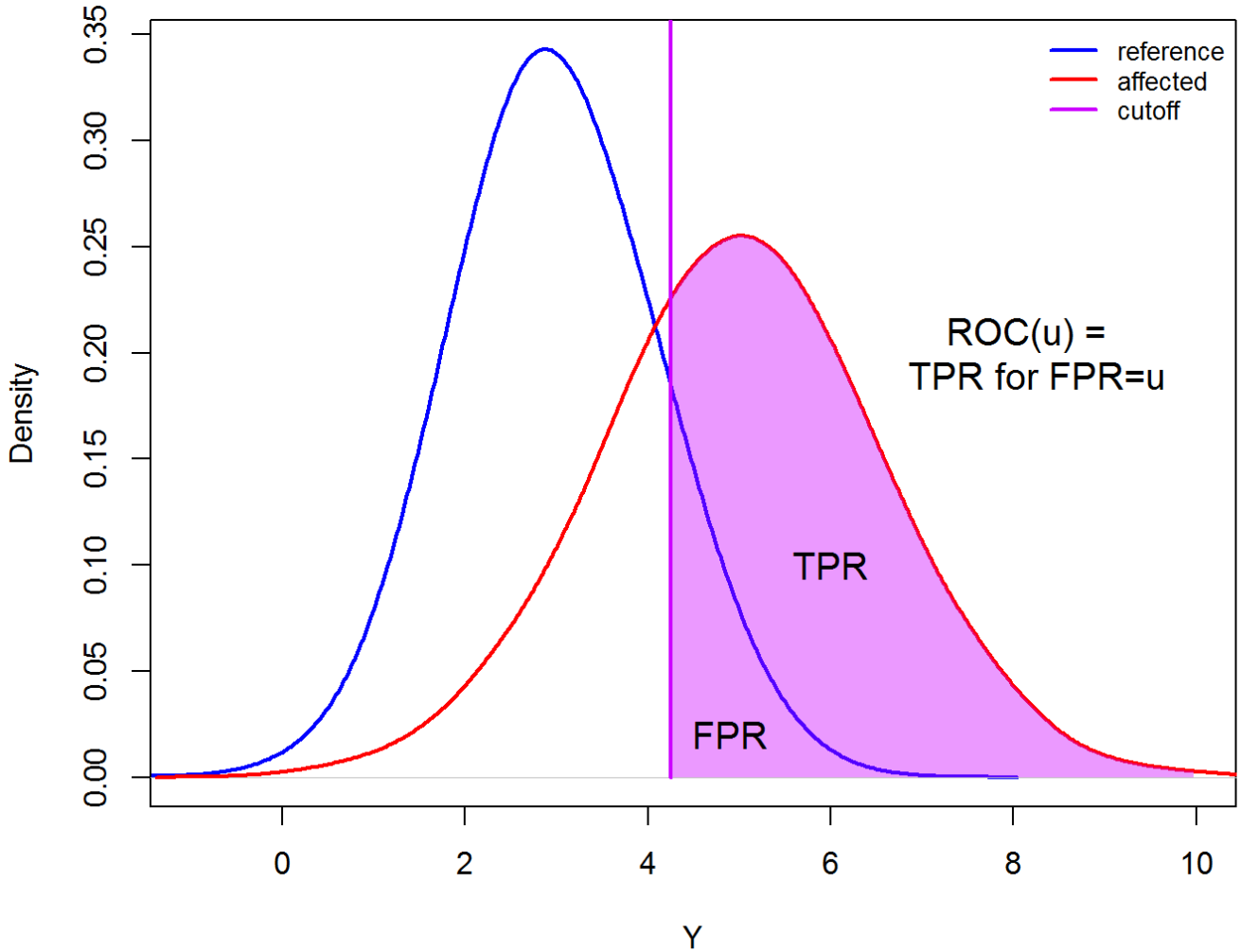
Quick Review



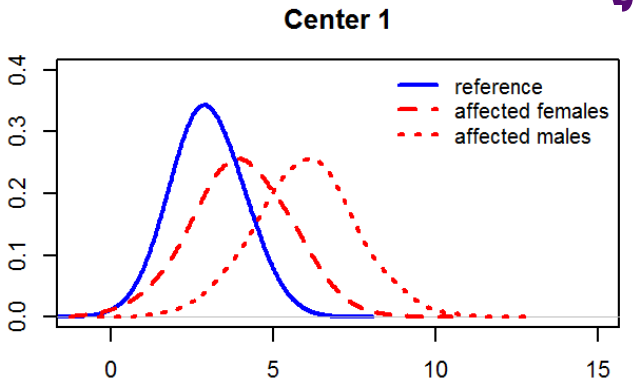
Quick Review

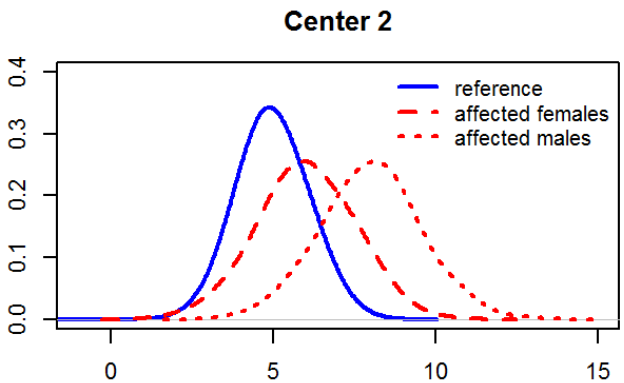


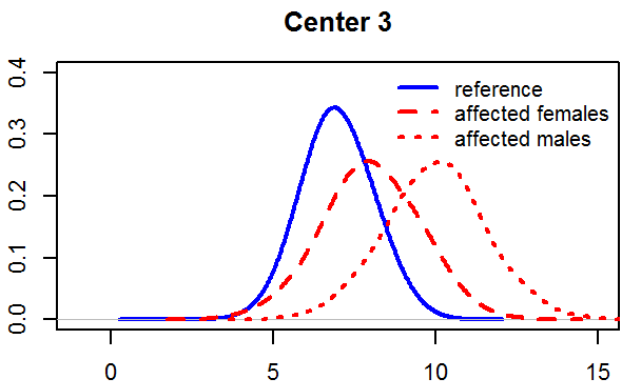
Quick Review



Quick Review

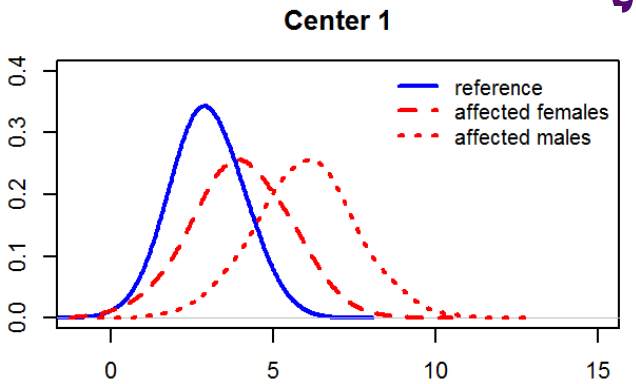


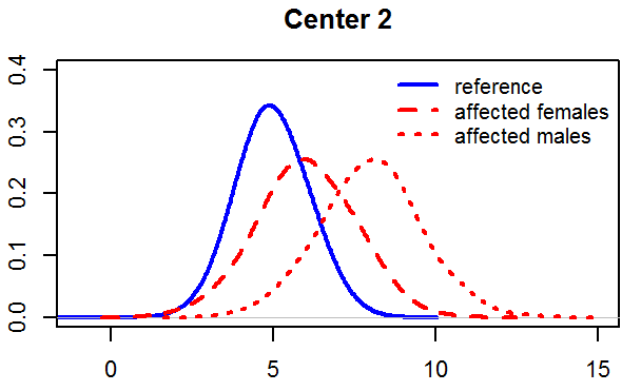


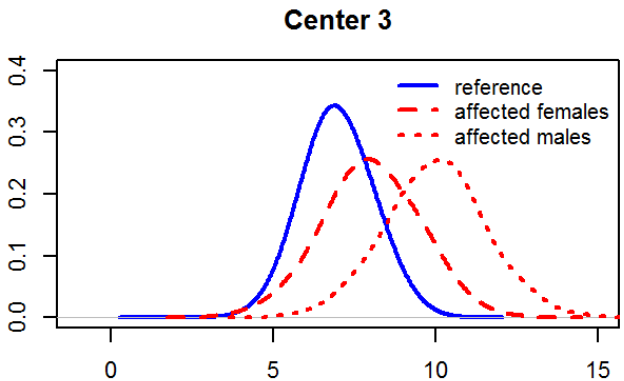


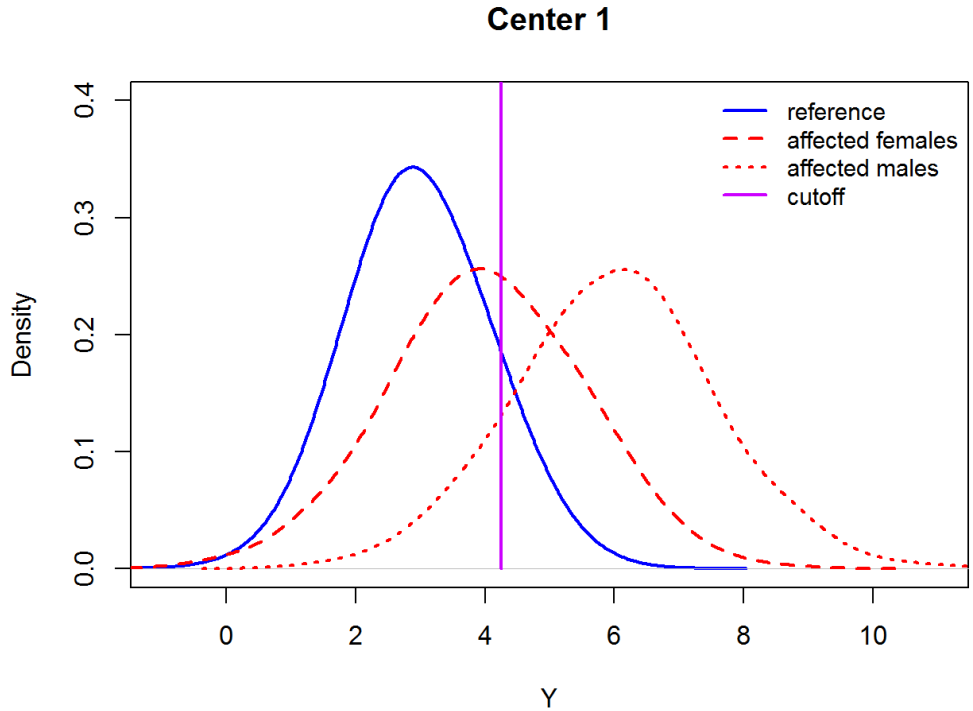
$\Rightarrow Z_{\overline{D}} = \text{center},\ Z_D = \text{sex}$

Quick Review



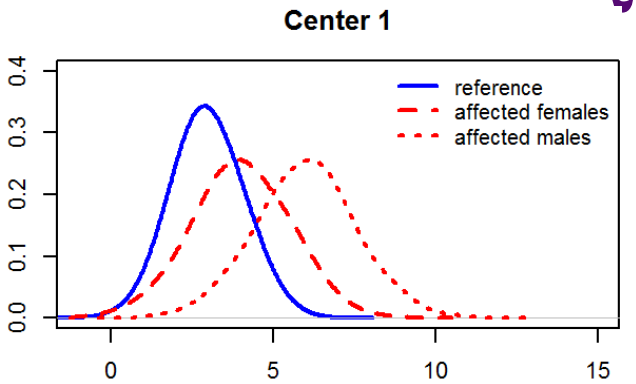


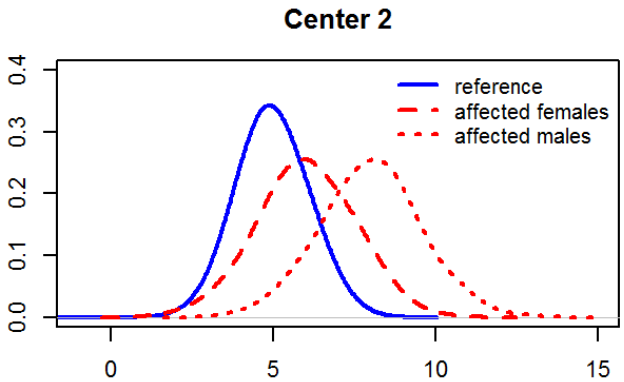


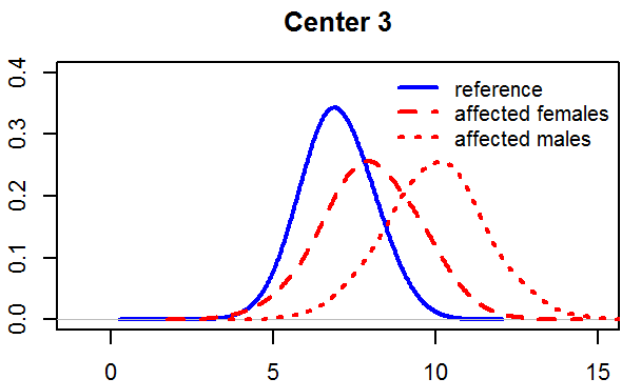


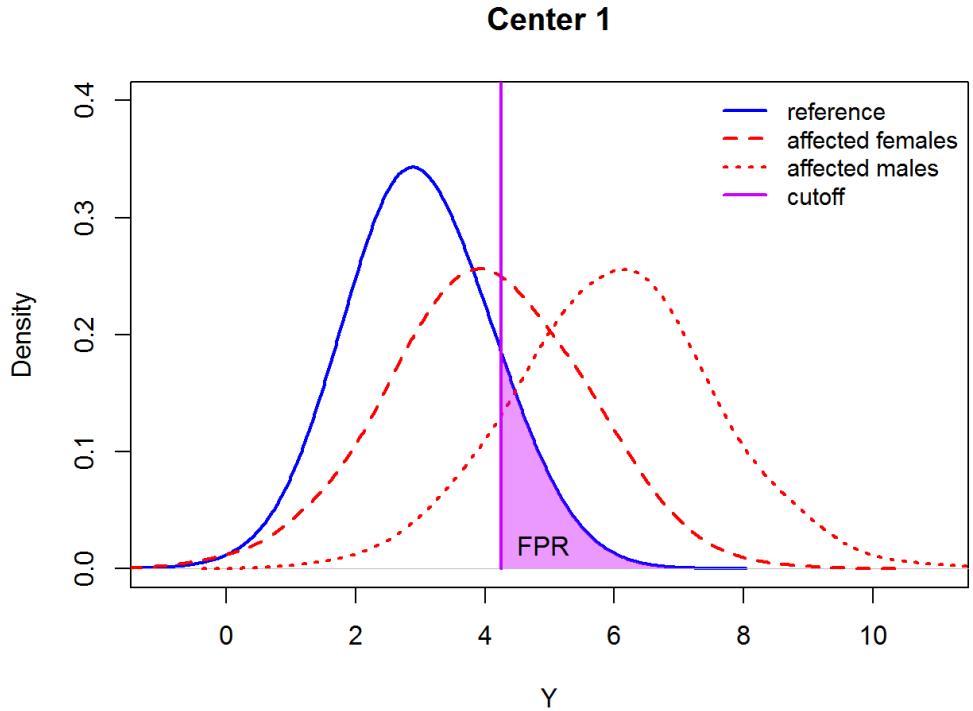
$\Rightarrow Z_{\overline{D}} = \text{center},\ Z_D = \text{sex}$

Quick Review



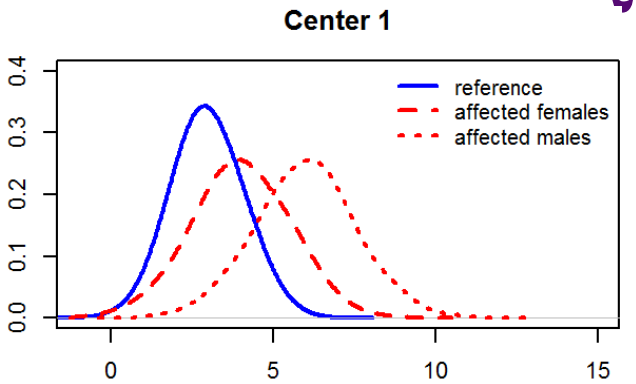


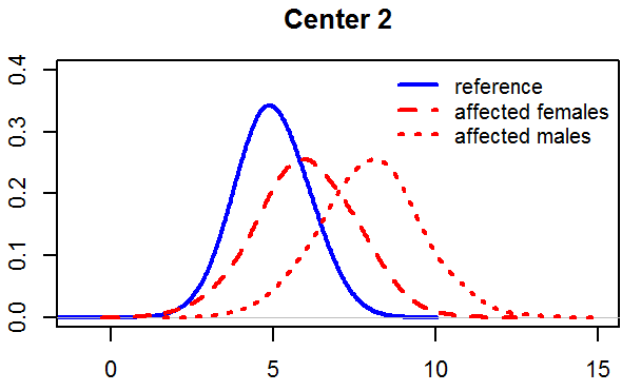


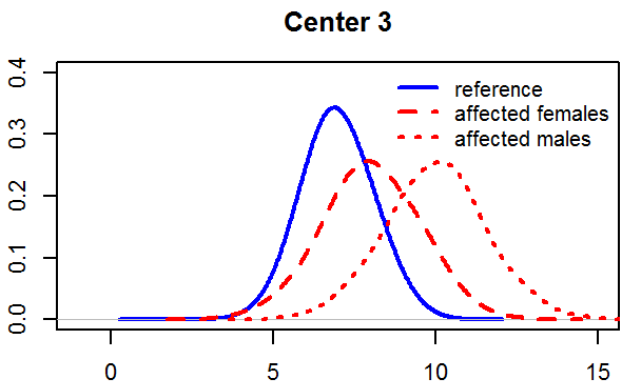


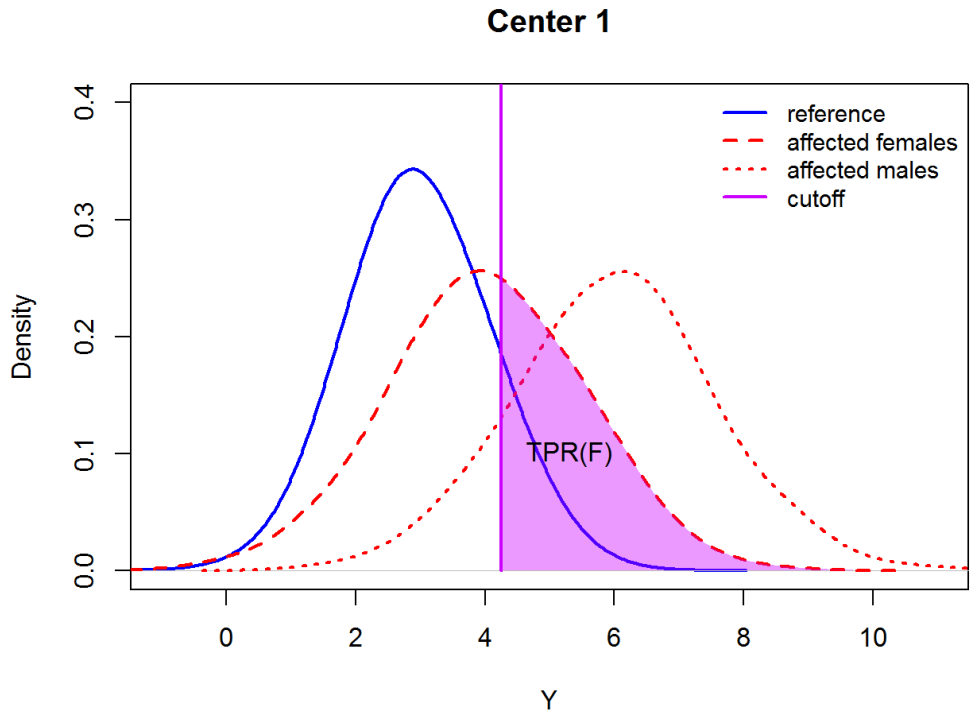
$\Rightarrow Z_{\overline{D}} = \text{center},\ Z_D = \text{sex}$

Quick Review



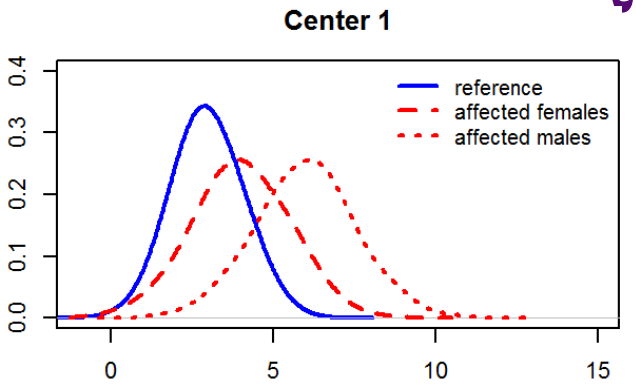


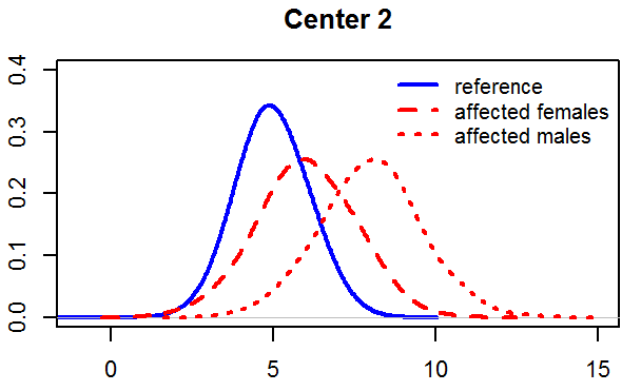


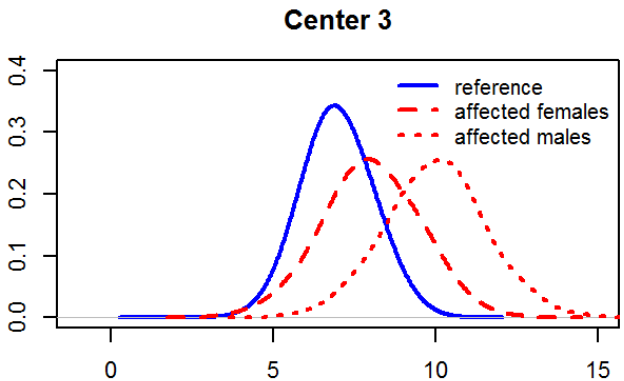


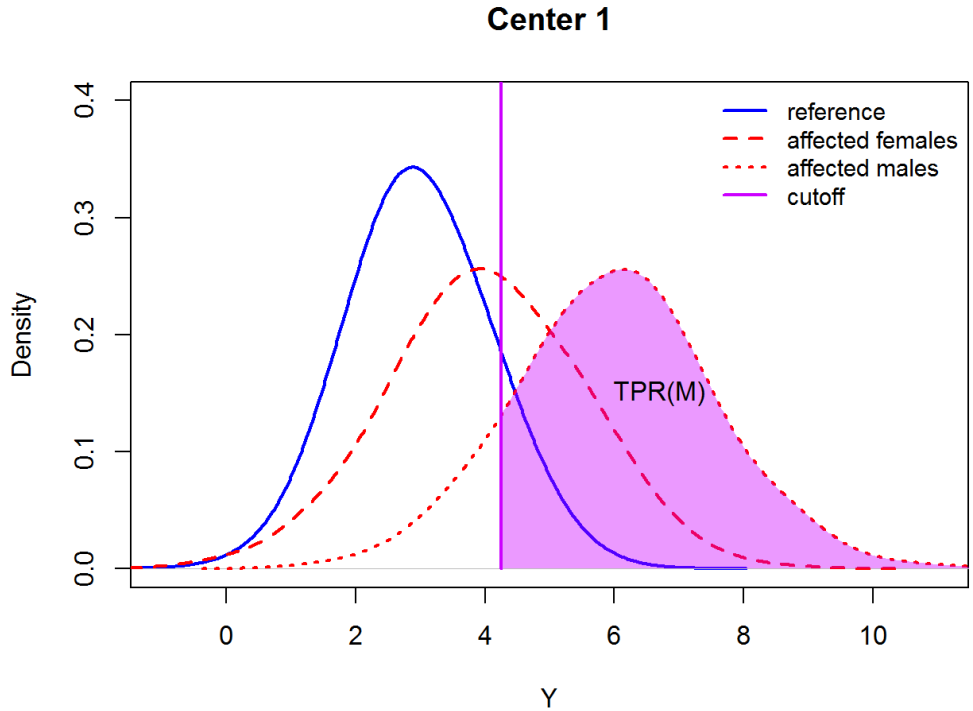
$\Rightarrow Z_{\overline{D}} = \text{center},\ Z_D = \text{sex}$

Quick Review



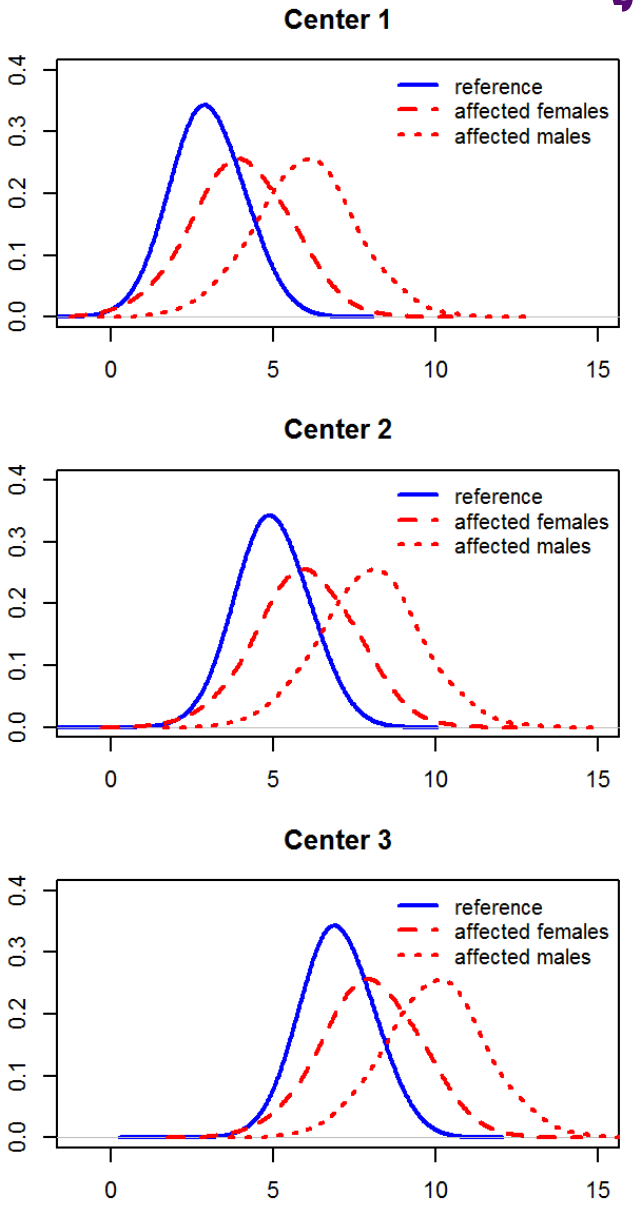


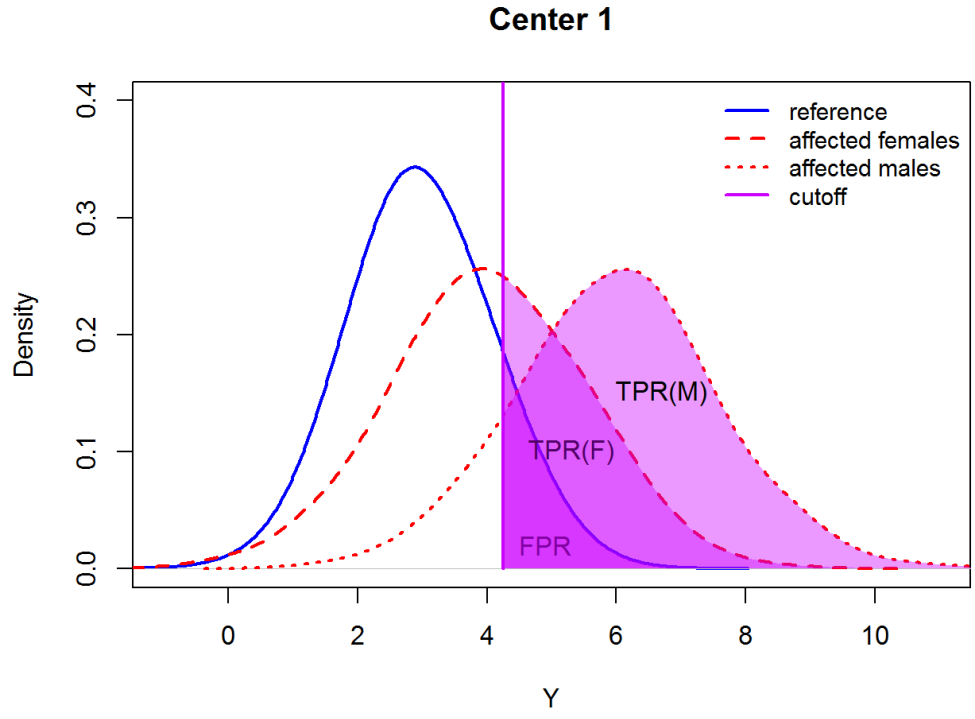




$\Rightarrow Z_{\overline{D}} = \text{center},\ Z_D = \text{sex}$

Quick Review





$\Rightarrow Z_{\overline{D}} = \text{center},\ Z_D = \text{sex}$

Definitions & Derivations

- Definitions
  - Placement values: $U_{D_i} = 1 - F_{\overline{D}}(Y_{D_i})$ for the i^{th} diseased subject
    If $Z_{\overline{D}}$ affects the distribution of Y in the reference population, $U_{D_i} = 1 - F_{\overline{D},Z_{\overline{D}}}(Y_{D_i})$
  - ROC curve: $ROC(u) = P(Y_D \geq F_{\overline{D}}^{-1}(1-u)) =$ (TPR at FPR=u)

- Relationship between ROC and placement values

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

Definitions & Derivations

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

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

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

$$ROC_{Z_{\overline{D}}, Z_D}(u) = g(\boldsymbol{\beta}^T Z_D + H_{\boldsymbol{\alpha}}(u))$$

Algorithm

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

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

Then, for $u \in [a, b]$, the log likelihood is

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

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

Algorithm

Estimating $F_{\overline{D},Z_{\overline{D}}}$

- Pepe and Cai advise estimating $F_{\overline{D},Z_{\overline{D}}}$ nonparametrically if $Z_{\overline{D}}$ 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, $\hat{U}_{D_i}$, are substituted into $\ell(\boldsymbol{\theta})$, yielding a pseudo-log-likelihood, which is maximized to estimate $\boldsymbol{\theta}$.

Simulations

Set-up

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

Induced ROC curve:

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

Simulations

Despite their recommendations, Pepe and Cai estimated placement values parametrically. (!) Note that here

$$Z_{\overline{D}} = Z_2 \text{ and } Z_D = (Z_1, Z_2).$$

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

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

Simulations

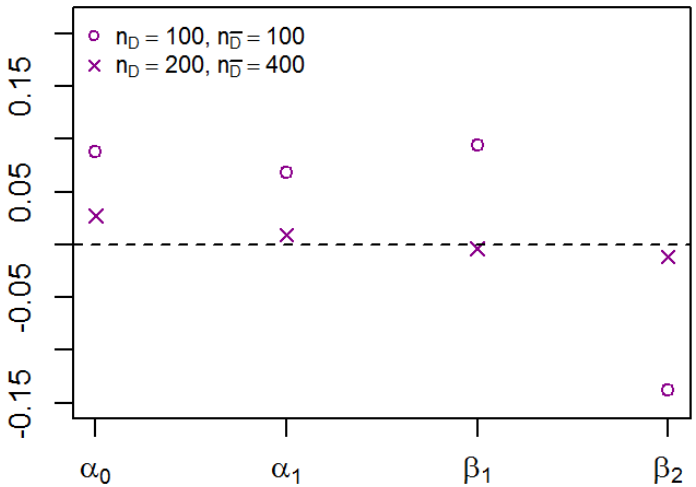
Two sets of simulations (1000 simulations):

1. Pepe and Cai method only
   - Bias
   - Empirical SE
   - Mean estimated SE (from 500 bootstrap samples)
   - 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
   - Considered $[a, b] = [0.01, 0.99]$ and $[a, b] = [0.01, 0.50]$

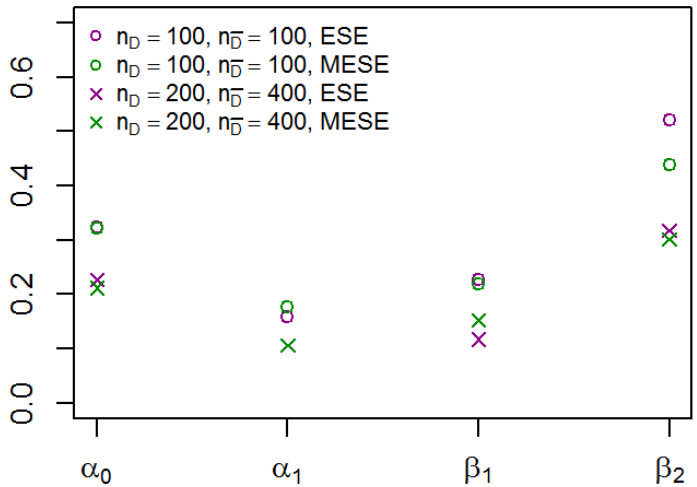
Simulations: Pepe and Cai

Results for a = 0.01, b = 0.99

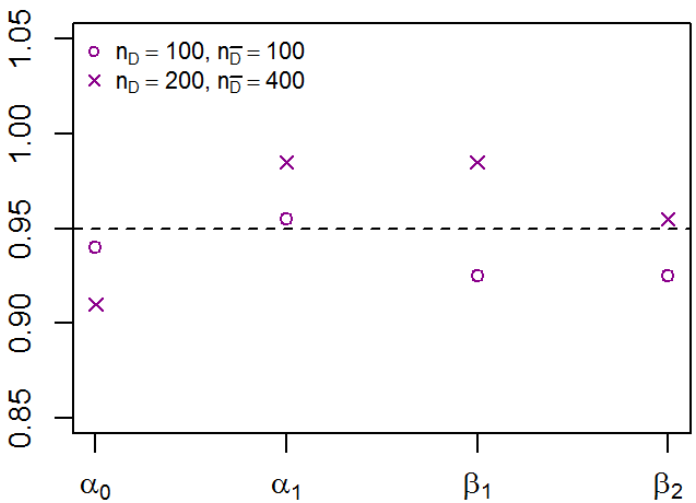
Percent Bias



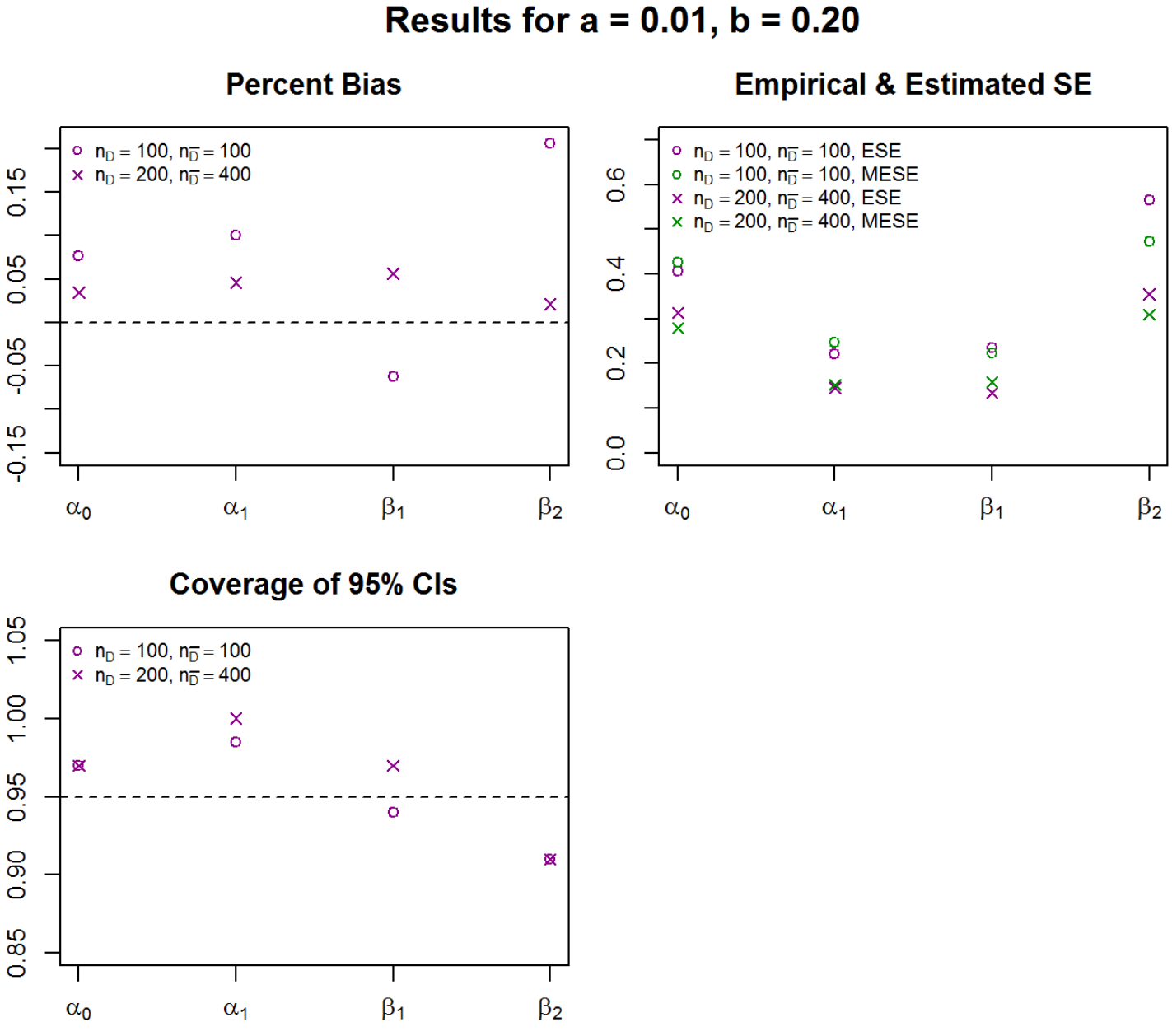
Empirical & Estimated SE



Coverage of 95% CIs



Simulations: Pepe and Cai



Next Steps

- Simulations with competing method
- Real data example
- Impact of nonlinear relationship between $Y_{\overline{D}}$ and Z_2 and/or impact of non-constant variance

Competing Method: Alonzo and Pepe

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

1. Choose a set of values in $[a, b]$ denoted by $T = \{u_1, ..., u_{n_T}\} = \{1 - j/n_{\overline{D}};\ \ j = 1, ..., n_{\overline{D}} - 1\}$.
2. Then for each diseased subject i, the n_T binary variables B_{ui} are calculated:

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

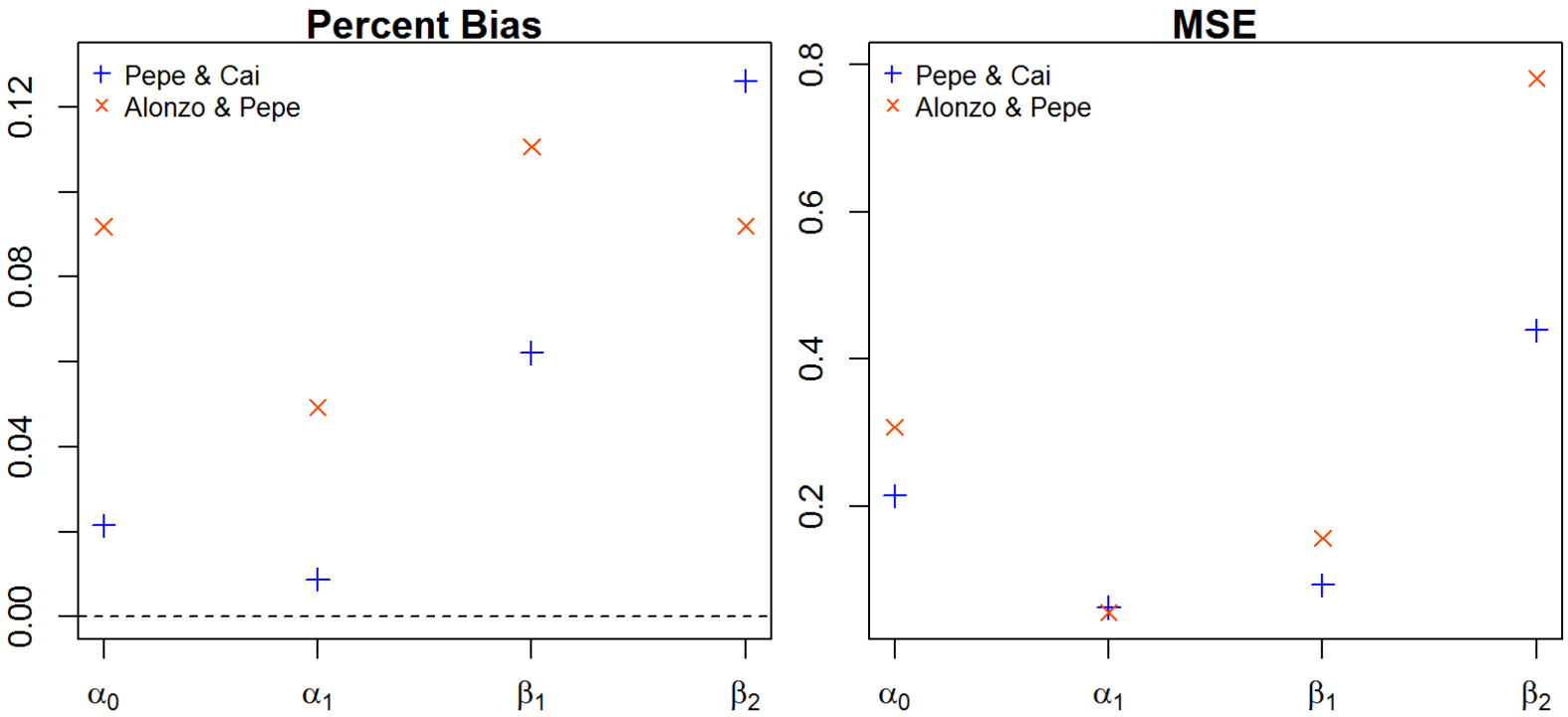
3. The binary generalized linear regression model

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

is fit using standard techniques.

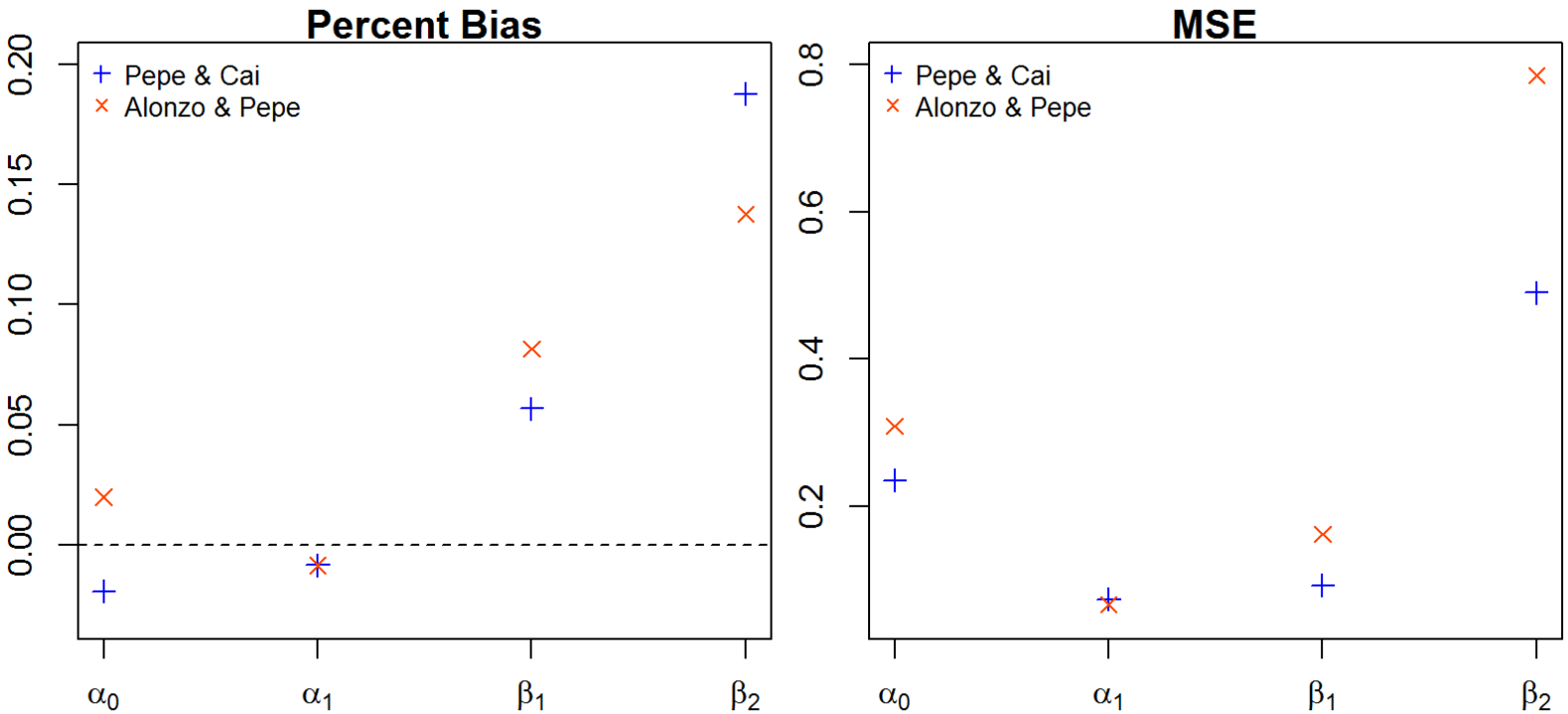
Simulations: Comparing Methods

- $n_D = n_{\overline{D}} = 50$
- a $= 0.01$, b $= 0.99$
- $\alpha_0 = 1, \alpha_1 = 1, \beta_1 = 0.5, \beta_2 = 0.7$



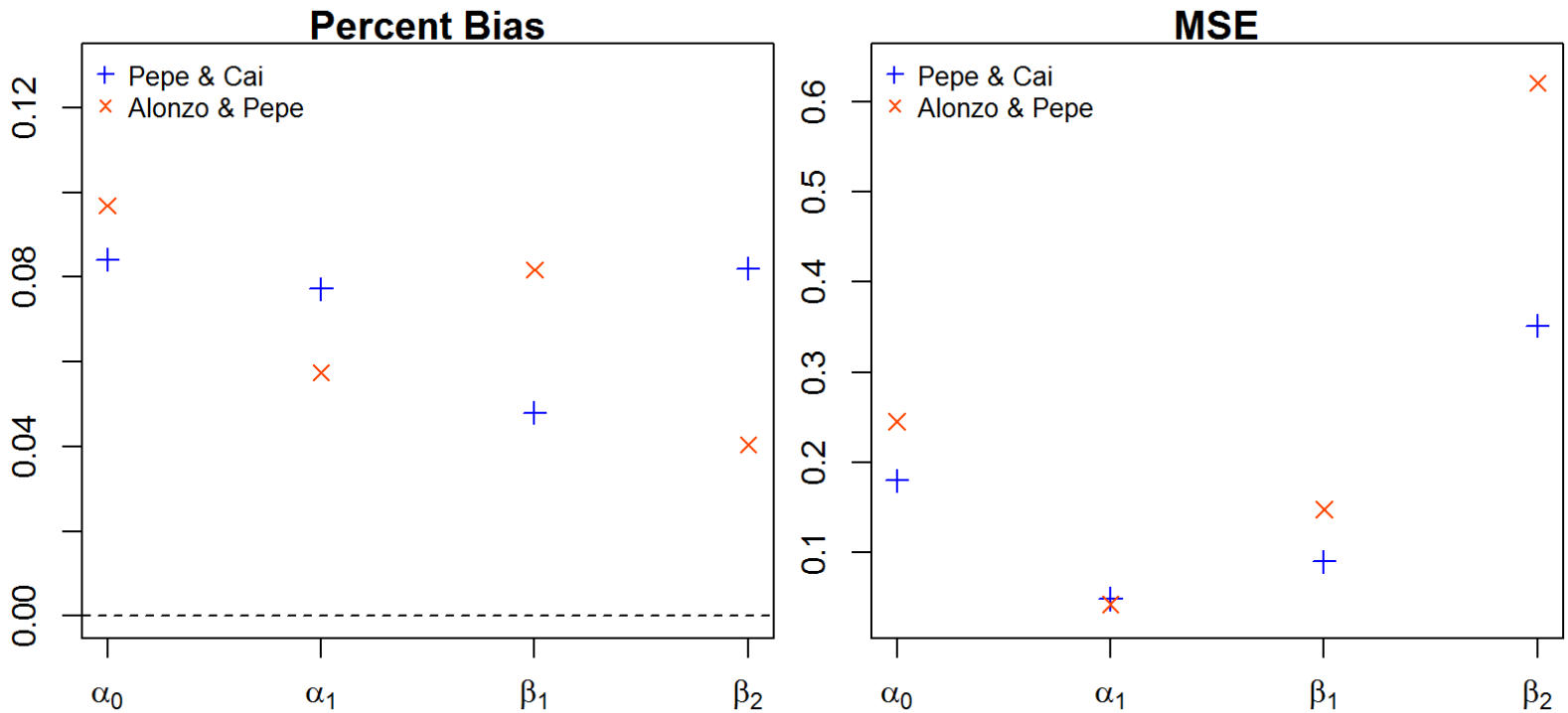
Simulations: Comparing Methods

- $n_D = n_{\overline{D}} = 50$
- a $= 0.01$, b $= 0.50$
- $\alpha_0 = 1, \alpha_1 = 1, \beta_1 = 0.5, \beta_2 = 0.7$



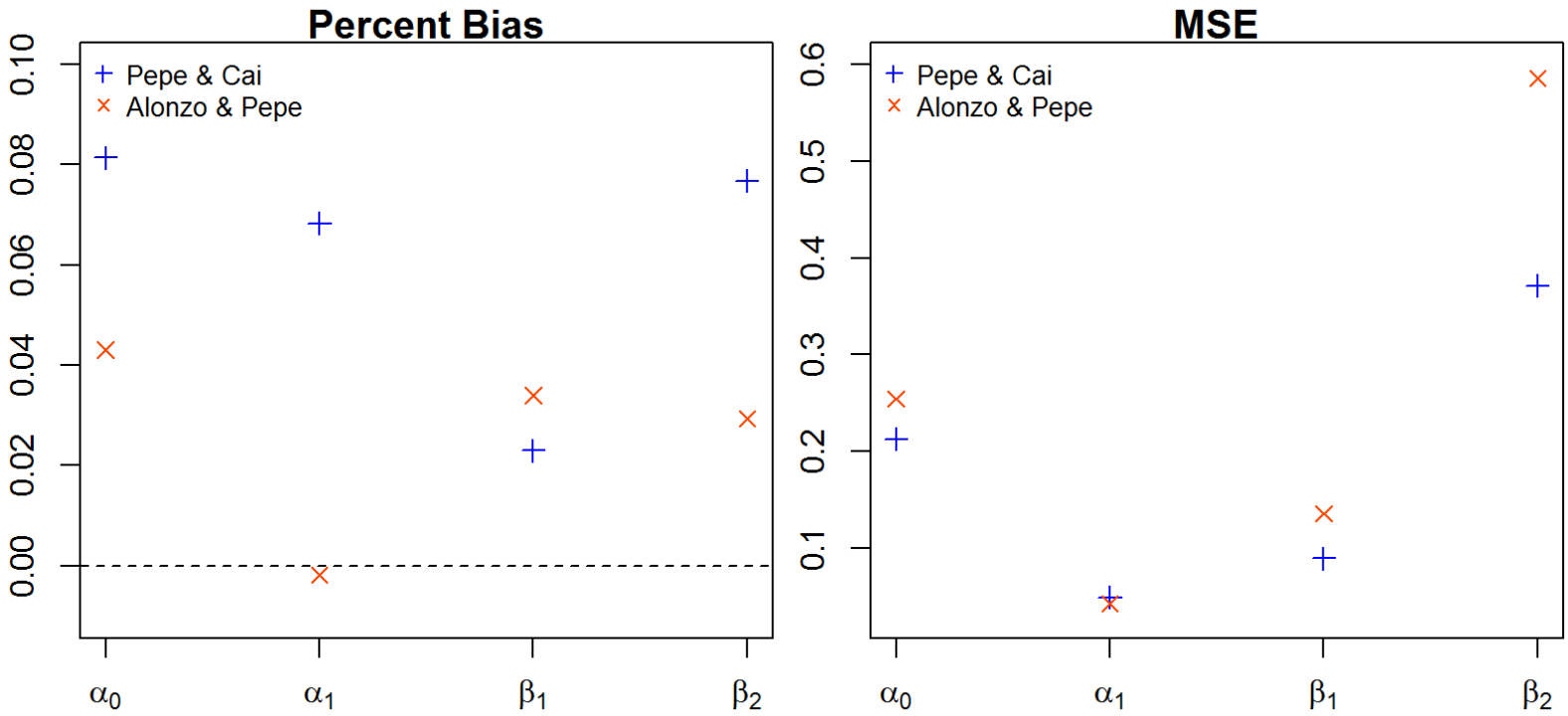
Simulations: Comparing Methods

- $n_D = 50, n_{\overline{D}} = 100$
- a $= 0.01$, b $= 0.99$
- $\alpha_0 = 1, \alpha_1 = 1, \beta_1 = 0.5, \beta_2 = 0.7$



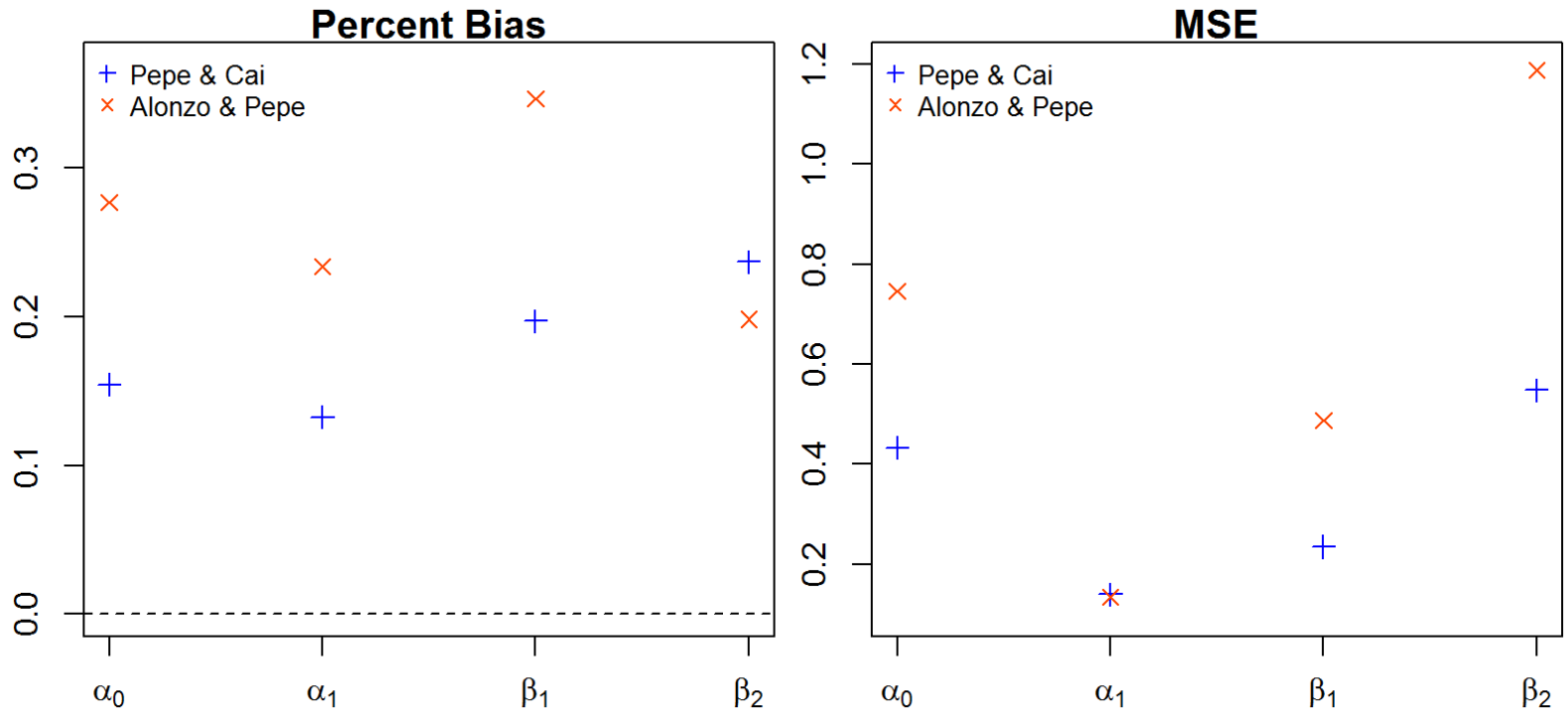
Simulations: Comparing Methods

- $n_D = 50, n_{\overline{D}} = 100$
- a $= 0.01$, b $= 0.50$
- $\alpha_0 = 1, \alpha_1 = 1, \beta_1 = 0.5, \beta_2 = 0.7$



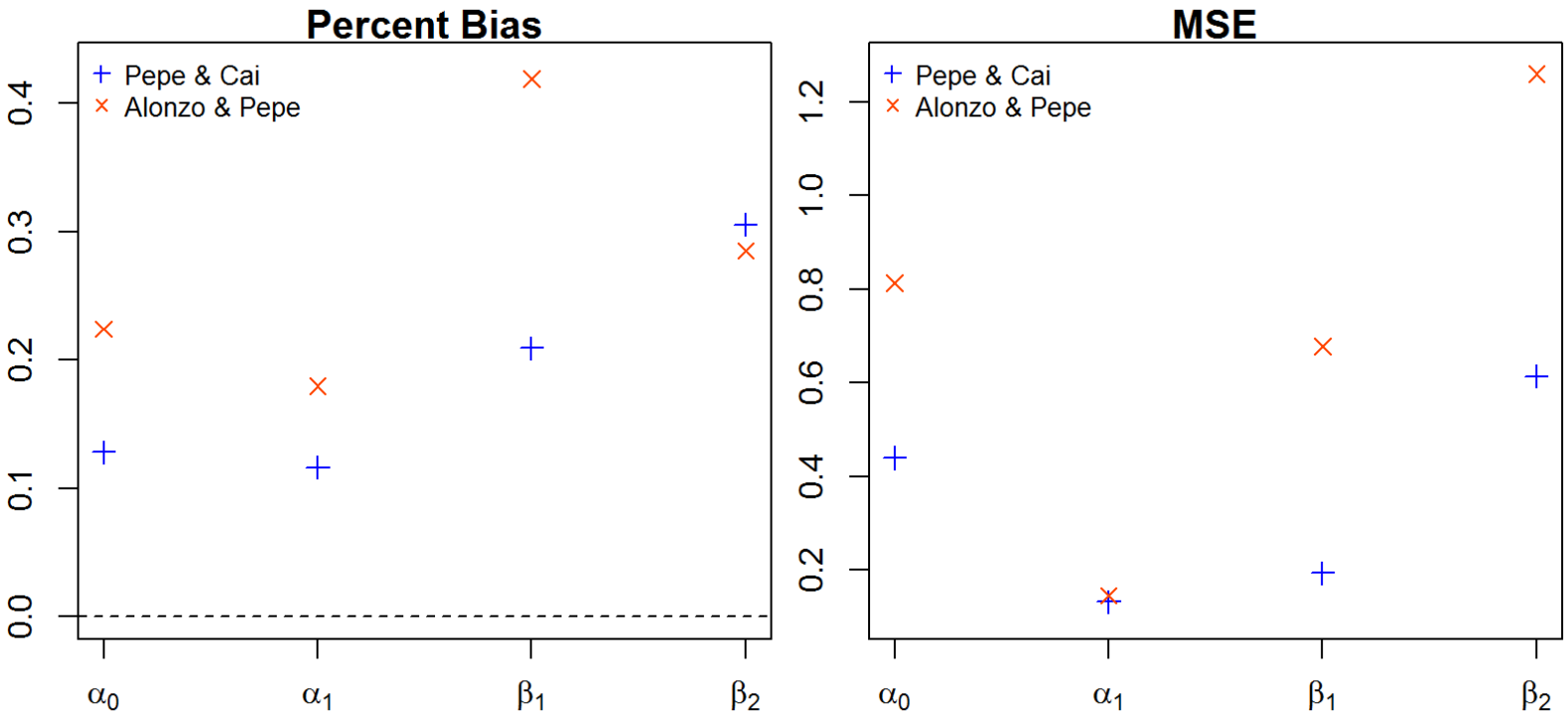
Simulations: Comparing Methods

- $n_D = n_{\overline{D}} = 50$
- a $= 0.01$, b $= 0.99$
- $\alpha_0 = 1.5, \alpha_1 = 0.9, \beta_1 = 0.5, \beta_2 = 0.7$



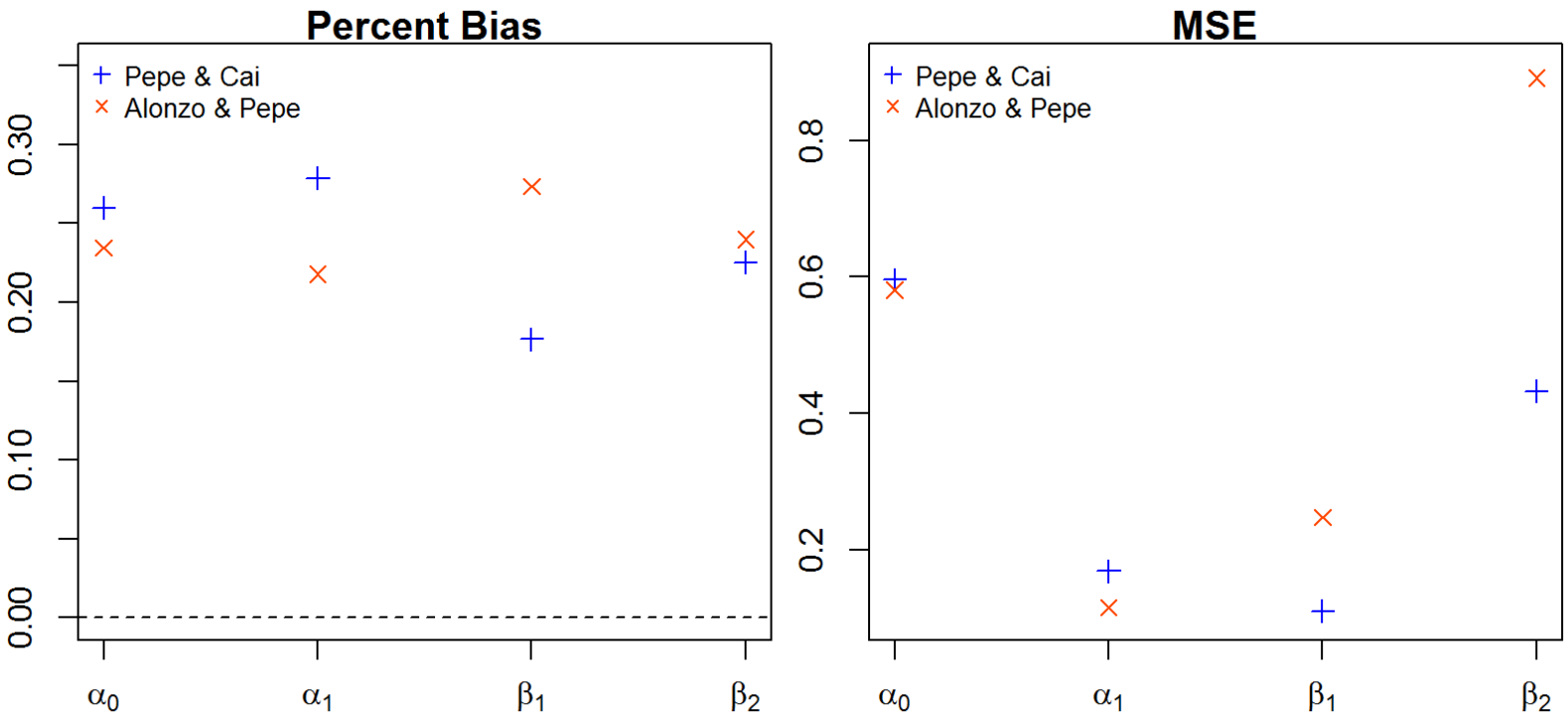
Simulations: Comparing Methods

- $n_D = n_{\overline{D}} = 50$
- a $= 0.01$, b $= 0.50$
- $\alpha_0 = 1.5, \alpha_1 = 0.9, \beta_1 = 0.5, \beta_2 = 0.7$



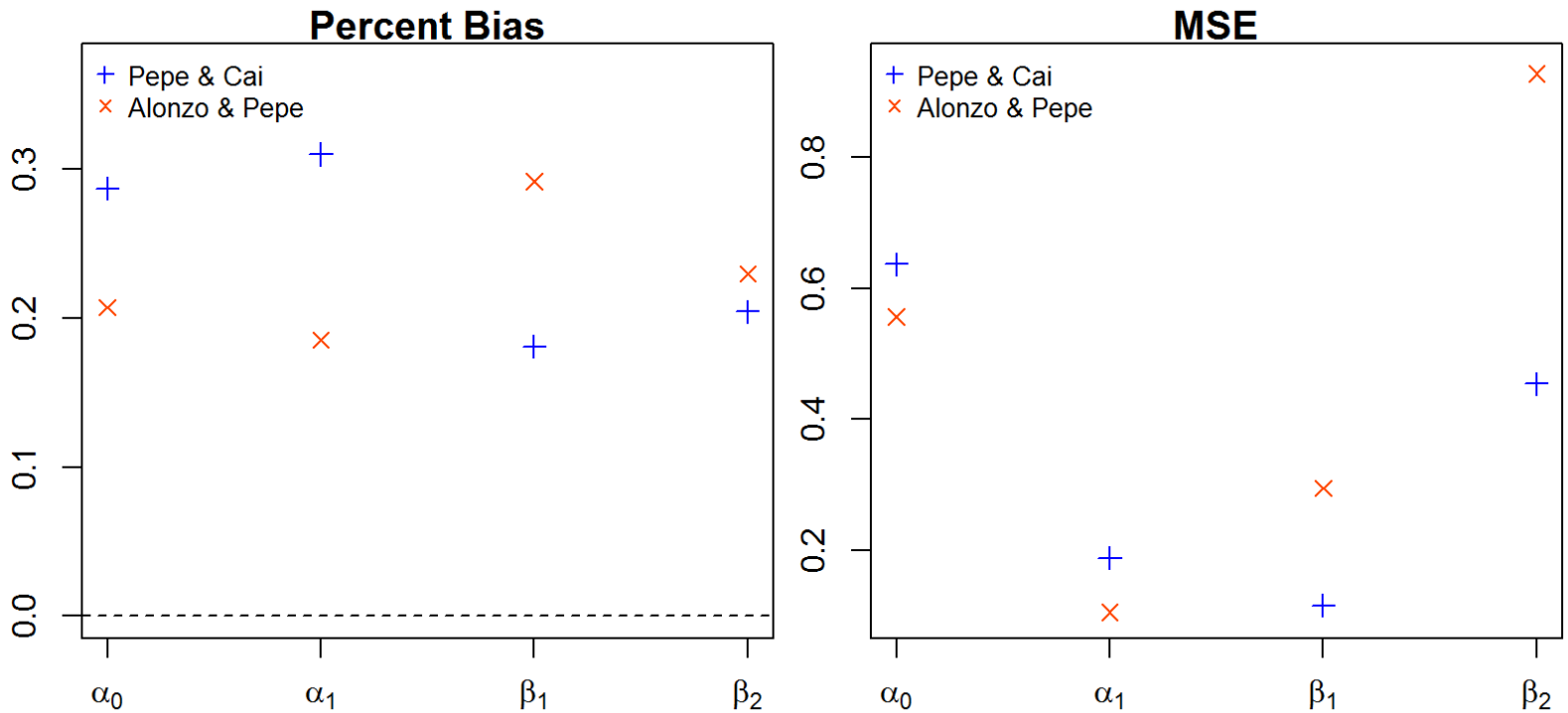
Simulations: Comparing Methods

- $n_D = 50, n_{\overline{D}} = 100$
- a $= 0.01$, b $= 0.99$
- $\alpha_0 = 1.5, \alpha_1 = 0.9, \beta_1 = 0.5, \beta_2 = 0.7$



Simulations: Comparing Methods

- $n_D = 50, n_{\overline{D}} = 100$
- a $= 0.01$, b $= 0.50$
- $\alpha_0 = 1.5, \alpha_1 = 0.9, \beta_1 = 0.5, \beta_2 = 0.7$



Next Steps

- Real data example
- Impact of nonlinear relationship between $Y_{\overline{D}}$ and Z_2 and/or impact of non-constant variance