Covariate adjusted measures of diagnostic accuracy based on pooled biomarkers

Christopher McMahan

Collaborators: Alexander McLain, Colin Gallagher, and Enrique Schisterman

October 13, 2016

Biological marker (or biomarker) evaluation

- The motivation behind evaluating new biomarkers:
  - Identify new markers that can be used to asses exposures
  - Identify new markers for disease detection
- In 2011, 70% of the original articles in *Clinical Chemistry*, were focused on biomarker evaluation; Boyd et al. (2012)
  - HIV; Kanekar (2010)
  - Cancer; Borges et al. (2013)
  - Cardiovascular disease; Sabatine et al. (2012)

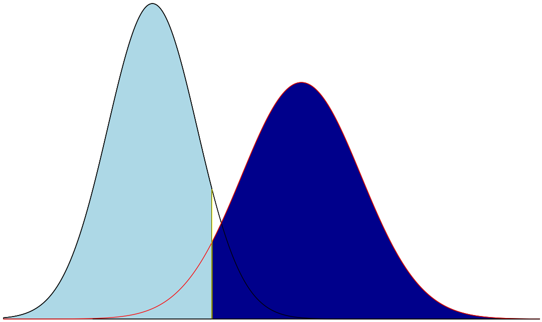
    ⋮
- This area of epidemiological research is often limited due to the cost associated with measuring biomarker levels
  - Caudill (2012) reported a cost of $1400 per specimen to obtain a single analytical measurement of 61 polychlorinated and 13 polybrominated compounds
- If you know me, you would know how I would solve this problem, more on this shortly

- Several common measures:
  - Receiver operating characteristic (ROC) curve
  - Area under the ROC curve (AUC)
  - Youden Index (YI)
- Let $f_{\mathcal{C}^-}$ and $f_{\mathcal{C}^+}$ denote the probability distribution functions for the biomarker levels associated with cases and controls, respectively
- Consider a test that diagnoses a subject as positive if their biomarker level is above a threshold t

$$\text{Sensitivity:} \qquad S_e(t) = P(\text{test} + |\text{truly}+) = \int_t^\infty f_{\mathcal{C}^+}(c)dc$$

$$\text{Specificity:} \qquad S_p(t) = P(\text{test} - |\text{truly}-) = \int_\infty^t f_{\mathcal{C}^-}(c)dc$$

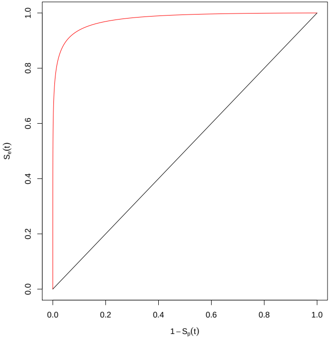
Measures of discriminatory ability



- $f_{\mathcal{C}^+}$ ($f_{\mathcal{C}^-}$) denoted by the red (black) curve
- t denoted by the yellow line
- $S_e(t)$ ($S_p(t)$) denoted by the dark (light) blue shaded region

Receiver operating characteristic (ROC) curve

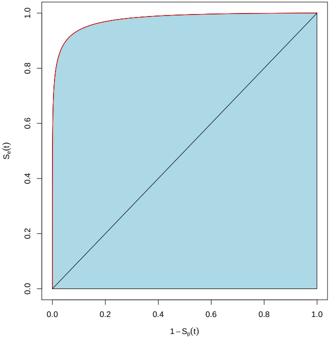
Construction: Plot $S_e(t)$ versus $1 - S_p(t)$, for all t



- If the ROC curve (red) is “far" from the chance line (black) then the biomarker is a good discriminator
- If the ROC curve (red) is “close" to the chance line (black) then the biomarker is not a useful discriminator

Area under the ROC curve (AUC)

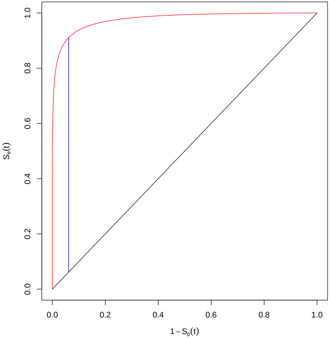
Calculation: $\text{AUC} = P(\mathcal{C}^+ > \mathcal{C}^-)$, where $\mathcal{C}^+ \sim f_{\mathcal{C}^+}$, $\mathcal{C}^- \sim f_{\mathcal{C}^-}$



- If AUC≈ 1, then the biomarker is a good discriminator
- If AUC≈ 0.5, then the biomarker is not a useful discriminator

Youden index (YI): Youden (1950)

Calculation: $\text{YI} = \sup_{t\in\mathbb{R}}\{S_e(t) + S_p(t) - 1\}$



- YI is the maximum vertical distance (blue) between the ROC curve (red) and the chance line (black)
- If YI$\approx$ 1(0), then the biomarker is an (in)effective discriminator

Pooled biomarker evaluation

- Pooling: A means to reduce testing cost
  - Physically combine several specimens into a pool and then measure the pool for the characteristic of interest
  - If one uses pools of size c, then N specimens can be assessed at the cost of $J = N/c$ measurements; i.e. at a drastic reduction in testing cost
- Dorfman (1943) originally proposed pool (group) testing
- Group testing has been used in many venues:
  - Infectious disease screening:
    - HIV, HBV, and HCV; Stramer et al. (2013)
    - Chlamydia and gonorrhea; Lindan et al. (2005)
  - Identifying lead compounds in drug discovery; Remlinger et al. (2006)
  - Screening for viral agents in the case of bioterrorism; Schmidt et al. (2005)
  - Detecting rare mutations in genetics; Gastwirth (2000)

Pooled biomarker evaluation

- Several authors have investigated the use of pooled assessments to evaluate the discriminatory ability of a biomarker of interest
  - Faraggi et al. (2003)
  - Liu and Schisterman (2003)
  - Mumford et al. (2006)
  - Bondell et al. (2007)
  - Vexler et al. (2008)
  - Malinovsky et al. (2012)
- Regretfully, all of these techniques have failed to acknowledge confounding factors (e.g., age, sex, gender, race, etc.)
- The focus of this work is to develop methods of estimating covariate dependent ROC curves, AUCs, and Youden indices based on pooled biomarker assessments

Models, notation, and assumptions

Control model:

$$Y_{ij-} = \mathbf{X}'_{ij-}\boldsymbol{\beta}_- + \epsilon_{ij-}, \text{ for } i = 1, ..., c_- \text{ and } j = 1, ..., J_-$$

Case model:

$$Y_{ij+} = \mathbf{X}'_{ij+}\boldsymbol{\beta}_+ + \epsilon_{ij+}, \text{ for } i = 1, ..., c_+ \text{ and } j = 1, ..., J_+$$

where,

- Y_{ij-} (Y_{ij+}) are the biomarker levels of the controls (cases)
- $\mathbf{X}_{ij-}$ ($\mathbf{X}_{ij+}$) is a p-dimensional vector of covariates
- $\boldsymbol{\beta}_-$ ($\boldsymbol{\beta}_+$) is a vector of regression parameters
- $\epsilon_{ij-} \stackrel{iid}{\sim} N(0, \sigma^2_-)$ and $\epsilon_{ij+} \stackrel{iid}{\sim} N(0, \sigma^2_+)$

Note: When pooled assessments are being made the individual level biomarker levels (i.e., Y_{ij-} and Y_{ij+}) are unobservable

Assumption: The aggregated, observed, pool response is the arithmetic average of the individuals biomarker levels

The models for the observed pooled assessments are
Control model:

$$Y_{j-} = \frac{1}{c_-}\sum_{i=1}^{c_-} Y_{ij-} = \overline{\mathbf{X}}'_{j-}\boldsymbol{\beta}_- + \epsilon_{j-}, \text{ for } j = 1, ..., J_-$$

Case model:

$$Y_{j+} = \frac{1}{c_+}\sum_{i=1}^{c_+} Y_{ij+} = \overline{\mathbf{X}}'_{j+}\boldsymbol{\beta}_+ + \epsilon_{j+}, \text{ for } j = 1, ..., J_+$$

where,

- $\overline{\mathbf{X}}_{j-} = c_-^{-1}\sum_{i=1}^{c_-}\mathbf{X}_{ij-}$ and $\overline{\mathbf{X}}_{j+} = c_+^{-1}\sum_{i=1}^{c_+}\mathbf{X}_{ij+}$
- $\epsilon_{j-} \overset{iid}{\sim} N(0, c_-^{-1}\sigma_-^2)$ and $\epsilon_{j+} \overset{iid}{\sim} N(0, c_+^{-1}\sigma_+^2)$

Parameter estimation

Model parameters are estimated as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_- &= (\overline{\mathbf{X}}'_-\overline{\mathbf{X}}_-)^{-1}\overline{\mathbf{X}}'_-\boldsymbol{Y}_- \\
\widehat{\boldsymbol{\beta}}_+ &= (\overline{\mathbf{X}}'_+\overline{\mathbf{X}}_+)^{-1}\overline{\mathbf{X}}'_+\boldsymbol{Y}_+ \\
\widehat{\sigma}^2_- &= c_-(J_- - p)^{-1}\boldsymbol{Y}'_-(\boldsymbol{I} - \boldsymbol{H}_-)\boldsymbol{Y}_- \\
\widehat{\sigma}^2_+ &= c_+(J_+ - p)^{-1}\boldsymbol{Y}'_+(\boldsymbol{I} - \boldsymbol{H}_+)\boldsymbol{Y}_+
\end{aligned}$$

where

- $\overline{\mathbf{X}}_- = (\overline{\mathbf{X}}_{1-}, ..., \overline{\mathbf{X}}_{J_-})'$ and $\overline{\mathbf{X}}_+ = (\overline{\mathbf{X}}_{1+}, ..., \overline{\mathbf{X}}_{J_++})'$
- $\boldsymbol{Y}^- = (Y_1^-, ..., Y_{J_--})'$ and $\boldsymbol{Y}^+ = (Y_1^+, ..., Y_{J_++})'$
- $\boldsymbol{H}_- = \overline{\mathbf{X}_-}(\overline{\mathbf{X}}'_-\overline{\mathbf{X}}_-)^{-1}\overline{\mathbf{X}}'_-$ and $\boldsymbol{H}_+ = \overline{\mathbf{X}}_+(\overline{\mathbf{X}}'_+\overline{\mathbf{X}}_+)^{-1}\overline{\mathbf{X}}'_+$
- $\boldsymbol{I}$ is the identity matrix

Parameter estimation

Under our modeling assumptions, it is easy to show that

$$\widehat{\boldsymbol{\beta}}_{-} \sim N\left(\boldsymbol{\beta}_{-}, c_{-}^{-1}\sigma_{-}^{2}(\overline{\mathbf{X}}_{-}'\overline{\mathbf{X}}_{-})^{-1}\right)$$
$$\widehat{\boldsymbol{\beta}}_{+} \sim N\left(\boldsymbol{\beta}_{+}, c_{+}^{-1}\sigma_{+}^{2}(\overline{\mathbf{X}}_{+}'\overline{\mathbf{X}}_{+})^{-1}\right)$$

and

$$\frac{(J_{-} - p)\widehat{\sigma}_{-}^{2}}{\sigma_{-}^{2}} \sim \chi^2_{J_{-}-p}$$
$$\frac{(J_{+} - p)\widehat{\sigma}_{+}^{2}}{\sigma_{+}^{2}} \sim \chi^2_{J_{+}-p}$$

Consequently, it is possible to conduct typical regression diagnostics, hypothesis tests, and inference

- Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_{+}, \boldsymbol{\beta}_{-}, \sigma_{+}^{2}, \sigma_{-}^{2})'$ denote the model parameters
- Let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}_{+}, \widehat{\boldsymbol{\beta}}_{-}, \widehat{\sigma}_{+}^{2}, \widehat{\sigma}_{-}^{2})'$ denote their estimates

Measure of discrimination

Covariate adjusted sensitivities and specificities:

$$S_e(\mathbf{X}, t, \boldsymbol{\theta}) = \Phi\left(\frac{\mathbf{X}'\boldsymbol{\beta}_+ - t}{\sigma_+}\right) \quad \text{and} \quad S_p(\mathbf{X}, t, \boldsymbol{\theta}) = \Phi\left(\frac{t - \mathbf{X}'\boldsymbol{\beta}_-}{\sigma_-}\right)$$

Covariate adjusted Youden index:

$$\mathrm{YI}(\mathbf{X}, \boldsymbol{\theta}) = \sup_{t \in \mathbb{R}} \left\{ \Phi\left(\frac{\mathbf{X}'\boldsymbol{\beta}_+ - t}{\sigma_+}\right) + \Phi\left(\frac{t - \mathbf{X}'\boldsymbol{\beta}_-}{\sigma_-}\right) - 1 \right\}$$

Covariate adjusted optimal cut point:

$$t_0(\mathbf{X}, \boldsymbol{\theta}) = \operatorname*{argmax}_{t \in \mathbb{R}} \left\{ \Phi\left(\frac{\mathbf{X}'\boldsymbol{\beta}_+ - t}{\sigma_+}\right) + \Phi\left(\frac{t - \mathbf{X}'\boldsymbol{\beta}_-}{\sigma_-}\right) - 1 \right\}$$

Covariate adjusted AUC:

$$\mathrm{AUC}(\mathbf{X}, \boldsymbol{\theta}) = \Phi\left(\frac{\mathbf{X}'\boldsymbol{\beta}_+ - \mathbf{X}'\boldsymbol{\beta}_-}{\sqrt{\sigma_+^2 + \sigma_-^2}}\right)$$

Estimation and inference

Estimates of the covariate adjusted Youden index, optimal cutpoint, and AUC can be obtained as

$$\mathrm{YI}(\mathbf{X}, \widehat{\boldsymbol{\theta}}), \quad t_0(\mathbf{X}, \widehat{\boldsymbol{\theta}}), \quad \mathrm{AUC}(\mathbf{X}, \widehat{\boldsymbol{\theta}})$$

Further, we establish that at a given predictor level $\mathbf{X}$

$$\begin{aligned}
\sqrt{J}\{\widehat{\mathrm{YI}}(\mathbf{X}, \widehat{\boldsymbol{\theta}}) - \mathrm{YI}(\mathbf{X}, \boldsymbol{\theta})\} &\xrightarrow{d} N(0, \Sigma_{\mathrm{YI}}) \\
\sqrt{J}\{\widehat{t}_0(\mathbf{X}, \widehat{\boldsymbol{\theta}}) - t_0(\mathbf{X}, \boldsymbol{\theta})\} &\xrightarrow{d} N(0, \Sigma_{t_0}) \\
\sqrt{J}\{\widehat{\mathrm{AUC}}(\mathbf{X}, \widehat{\boldsymbol{\theta}}) - \mathrm{AUC}(\mathbf{X}, \boldsymbol{\theta})\} &\xrightarrow{d} N(0, \Sigma_{\mathrm{AUC}})
\end{aligned}$$

- The above expressions assume that $J_- = J_+ = J$
- Closed form expressions (along with their finite sample estimators) of the asymptotic variances (i.e., Σ_{YI}, Σ_{t_0}, and Σ_{AUC}) were also obtained

To simultaneously assess a biomarker across the entire covariate space we derive asymptotic $100(1-\alpha)\%$ confidence bands for $\text{AUC}(\mathbf{X}, \boldsymbol{\theta})$; i.e., the sets $C(\mathbf{X})$ can be constructed such that

$$\text{pr}\left\{\text{AUC}(\mathbf{X}, \boldsymbol{\theta}) \in C(\mathbf{X}) \text{ for all } \mathbf{X}\right\} = 1 - \alpha.$$

Sets of this form can be constructed as

$$C(\mathbf{X}) = \left[\Phi\left(\frac{\mathbf{X}'(\widehat{\boldsymbol{\beta}}^{+} - \widehat{\boldsymbol{\beta}}^{-})}{\sqrt{\widehat{\sigma}_{+}^{2} + \widehat{\sigma}_{-}^{2}}} - \sqrt{\chi^{2}_{p,1-\alpha}}\sqrt{\widehat{\Sigma}_{\text{AUC}^{*}}}\right), \Phi\left(\frac{\mathbf{X}'(\widehat{\boldsymbol{\beta}}^{+} - \widehat{\boldsymbol{\beta}}^{-})}{\sqrt{\widehat{\sigma}_{+}^{2} + \widehat{\sigma}_{-}^{2}}} + \sqrt{\chi^{2}_{p,1-\alpha}}\sqrt{\widehat{\Sigma}_{\text{AUC}^{*}}}\right)\right],$$

where

- $\chi^{2}_{p,1-\alpha}$ denotes the $1-\alpha$th quantile of a chi-squared distribution with p degrees of freedom
- $\widehat{\Sigma}_{\text{AUC}^{*}}$ is an asymptotic variance estimator whose explicit form is provided in our manuscript

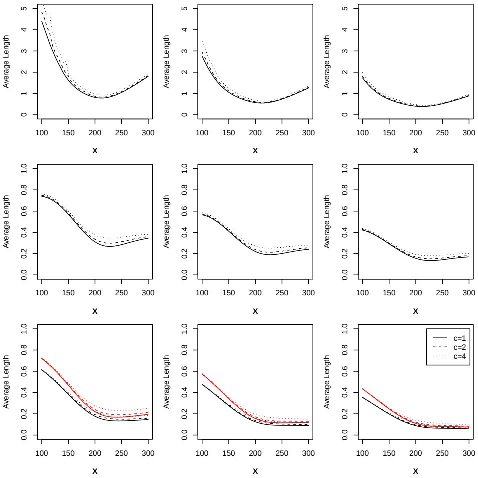
Simulation study

Simulation settings:

$$\textbf{Control model:} \quad Y_{k-} = \mathbf{X}'_{k-}\boldsymbol{\beta}_{-} + \epsilon_{k-} \quad \text{for } k = 1, \ldots., N,$$
$$\textbf{Case model:} \quad Y_{k+} = \mathbf{X}'_{k+}\boldsymbol{\beta}_{+} + \epsilon_{k+} \quad \text{for } k = 1, \ldots., N,$$

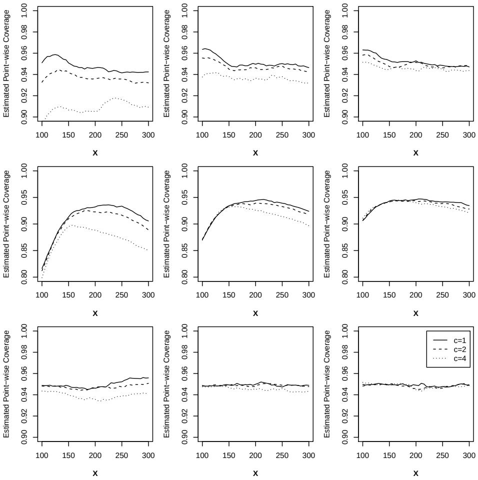
- ❏ $\mathbf{X}_{k+} = (1, X_{k1+})'$ and $X_{k1+} \sim N(225, 40^2)$
- ❏ $\mathbf{X}_{k-} = (1, X_{k1-})'$ and $X_{k1-} \sim N(205, 40^2)$
- ❏ $\epsilon_{k+} \sim N(0, 2.15^2)$, and $\epsilon_{k-} \sim N(0, 1.35^2)$
- ❏ $\boldsymbol{\beta}_{+} = (1.750, 0.015)'$ and $\boldsymbol{\beta}_{-} = (3.000, -0.005)'$
- ❏ Sample sizes: $N \in \{40, 80, 160\}$
- ❏ Pool sizes: $c_{-} = c_{+} = c \in \{1, 2, 4\}$
- ❏ Two pooling schemes: Random pooling (RP) and homogeneous pooling (HP)
- ❏ Replications: For each (c, N) combination and pooling scheme 10,000 data sets were generated and analyzed

Simulation study



- Top to bottom: $t_0(\mathbf{X}, \boldsymbol{\theta})$, $\mathrm{YI}(\mathbf{X}, \boldsymbol{\theta})$, and $\mathrm{AUC}(\mathbf{X}, \boldsymbol{\theta})$
- Left to right: N=40, 80, and 160

Simulation study

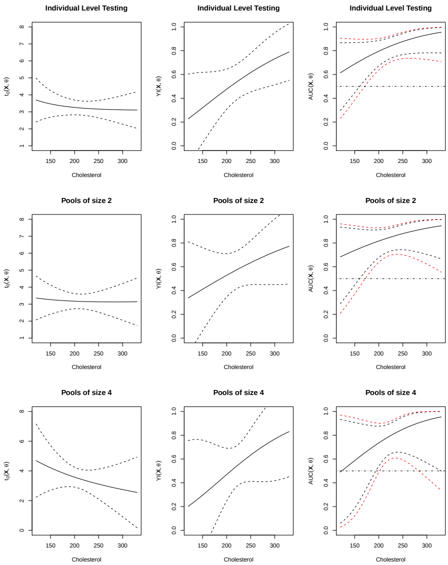


- Top to bottom: $t_0(\mathbf{X}, \boldsymbol{\theta})$, $\text{YI}(\mathbf{X}, \boldsymbol{\theta})$, and $\text{AUC}(\mathbf{X}, \boldsymbol{\theta})$
- Left to right: N=40, 80, and 160

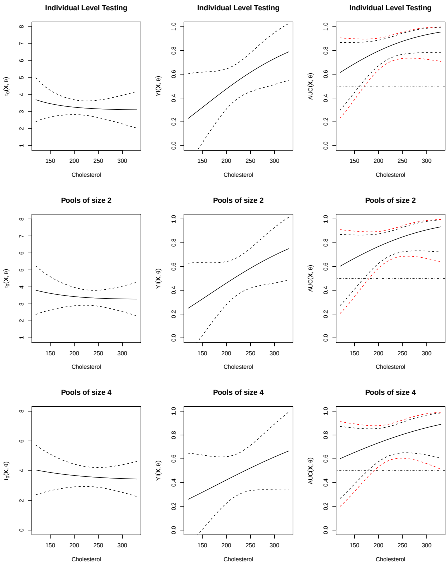
Data application

- Interleukin-6 (IL-6) is a pro-inflammatory cytokine that has been related to a host of biological functions, including coronary heart disease
- High levels of cholesterol are also associated with coronary heart disease
- This analysis considers 40 cases who had recently had a myocardial infarction (MI), and 40 controls
- Cholesterol and IL-6 were measured on all 80 subjects individually
- IL-6 was also assessed on pools of size two and four under RP
- For comparative purposes, we also consider artificially implementing HP
- A first order linear model was fit to the case and control data separately, using cholesterol as the only predictor variable

Results of data analysis: Observed data



Results of data analysis: Artificial HP



Discussion and future work

- Developed regression methodology for pooled biomarker measurements
- The proposed methodology allows one to estimate and perform inference about several common covariate dependent measures of discrimination; i.e., ROC, YI, AUC, and t_0
- Through additional simulation studies, we have discovered that our proposed techniques are relatively robust to departures from normality
- Future work includes, but is not limited to:
  - Extending the methodology proposed here to the class of generalized linear models
  - Develop nonparametric/semiparametric alternatives
  - Generalize to allow for the analysis of multiple biomarkers simultaneously
  - Account for common issues; e.g., measurement error, limits of detection, etc.