Reducing the Sample Size of Diagnostic-Biomarker-Validation Designs by a Bayesian Framework

L. Garcia Barrado[1] E. Coart[2] T. Burzykowski[1,2]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-Biostat)
[2] International Drug Development Institute (IDDI)

Outline

Outline

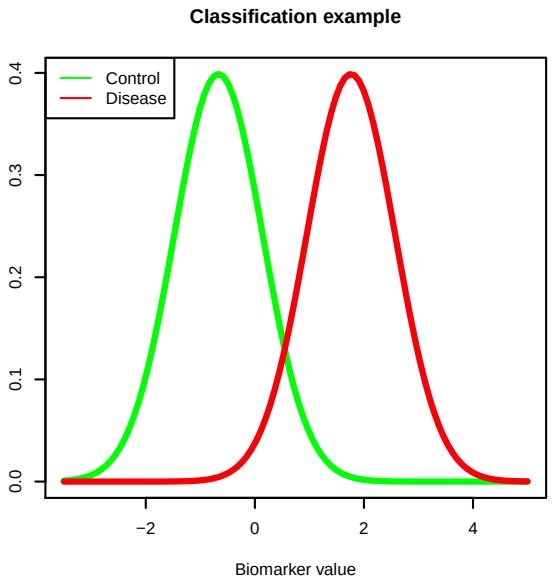
Problem setting

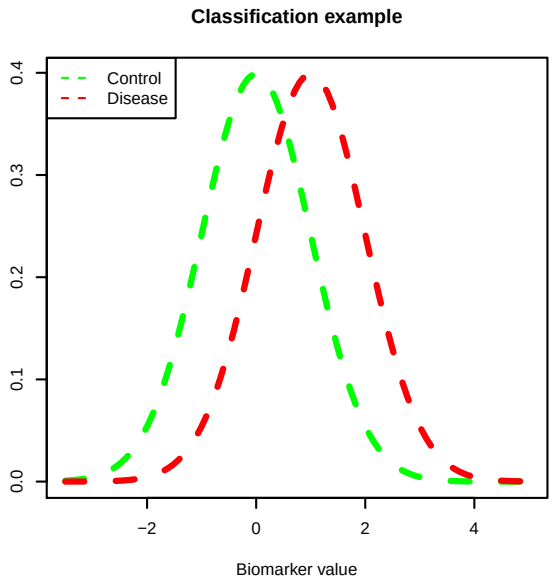
Develop a diagnostic biomarker-index and efficiently validate its accuracy

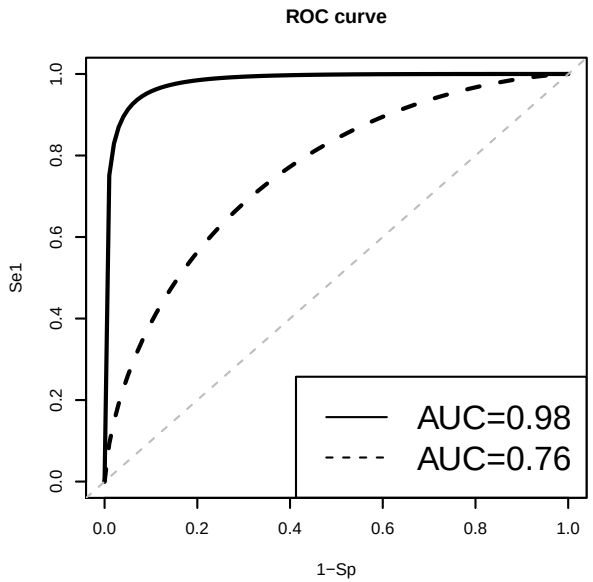
- Area under the Receiver Operating Characteristics (ROC) curve (AUC) as measure of accuracy
- Define index as linear combination of biomarkers maximizing AUC
- Incorporate information from development to validation stage

=> To this end a Bayesian framework will be proposed

Area under the Receiver Operating Characteristics curve







Data assumptions and notation

Underlying true biomarker distribution

- Mixture of two K-variate normal distributions by true disease status (D)
  - $\mathbf{Y}|_{D=0} \sim N_K(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$
  - $\mathbf{Y}|_{D=1} \sim N_K(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
- Reference test (T) is imperfect
  - Se: Unknown sensitivity of the reference test
  - Sp: Unknown specificity of the reference test
  - Conditionally on true disease status, misclassification independent of biomarker value
- θ: Unknown true prevalence of disease in the data set
- Assume for the rest of the presentation K=3

Definintion of biomarker-index

Linear combination maximizing AUC of the form*:

$$\mathbf{a'Y}|_{D=0} \sim N(\mathbf{a'}\boldsymbol{\mu_0}, \mathbf{a'}\boldsymbol{\Sigma_0}\mathbf{a})$$
$$\mathbf{a'Y}|_{D=1} \sim N(\mathbf{a'}\boldsymbol{\mu_1}, \mathbf{a'}\boldsymbol{\Sigma_1}\mathbf{a})$$

For which:

$$\mathbf{a'} \propto (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$$

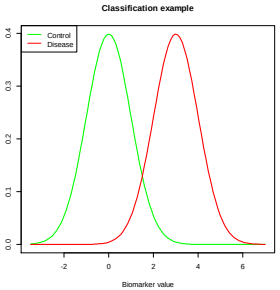
Area Under the ROC Curve:

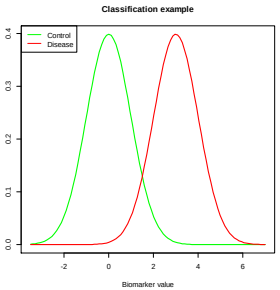
$$AUC_{Index} = \Phi\left\{\left[(\boldsymbol{\mu_1} - \boldsymbol{\mu}_0)'(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu}_0)\right]^{\frac{1}{2}}\right\}$$

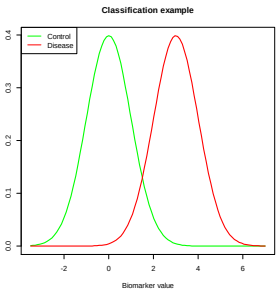
*Siu, J.Q., and Liu, J.S. (1993)

Development

Data

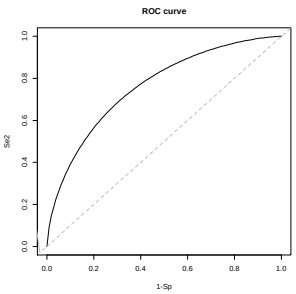






Results

$$\hat{a}_1$$
$$\hat{a}_2$$
$$\hat{a}_3$$



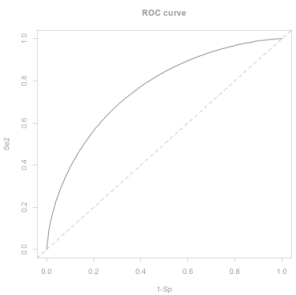
Development

Validation

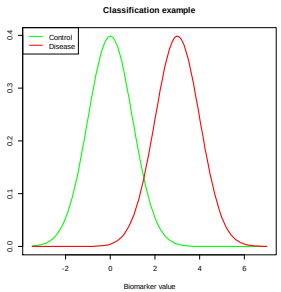
Data

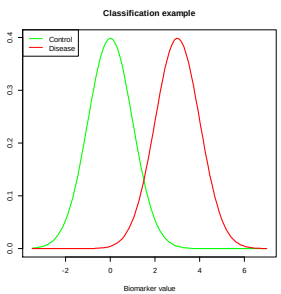
Results

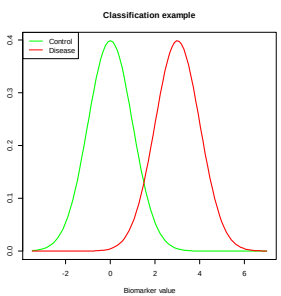
$$\hat{a}_1$$
$$\hat{a}_2$$
$$\hat{a}_3$$



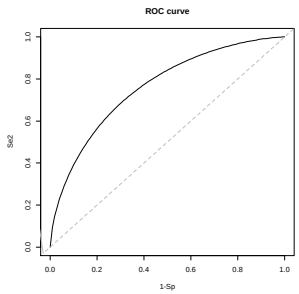
Data







Results



IDDI ADDING VALUE TO CLINICAL DATA
universiteit hasselt | I-BioStat | KU LEUVEN
Interuniversity Institute for Biostatistics and statistical Bioinformatics

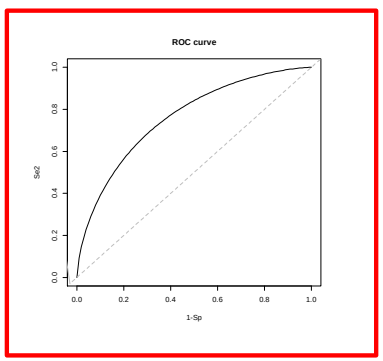
Development

Validation

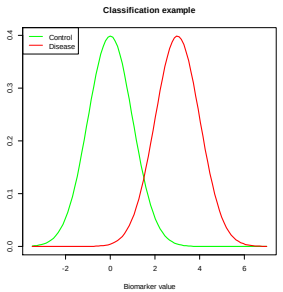
Data

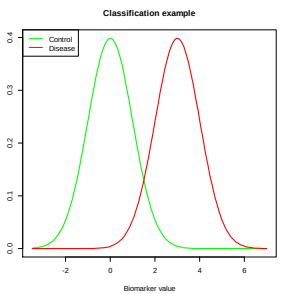
Results

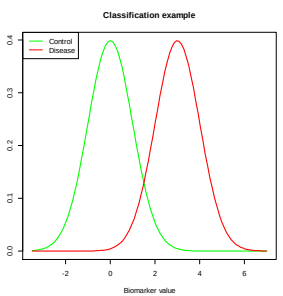
$$\hat{a}_1$$
$$\hat{a}_2$$
$$\hat{a}_3$$



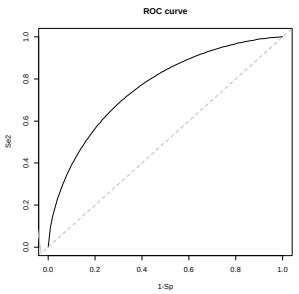
Data







Results



Outline

Bayesian latent-class mixture model

Full data likelihood

$$L(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \theta, Se, Sp | \mathbf{Y}, \mathbf{T}, \mathbf{D})$$

$$= \prod_{i=1}^{N} \left(\theta Se^{t_i} (1 - Se)^{(1-t_i)} \frac{1}{\sqrt{2\pi |\boldsymbol{\Sigma}_1|}} \times EXP \left\{ -\frac{1}{2} (\mathbf{Y}_i - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_1) \right\} \right)^{d_i}$$

$$\times \left((1 - \theta)(1 - Sp)^{t_i} Sp^{(1-t_i)} \frac{1}{\sqrt{2\pi |\boldsymbol{\Sigma}_0|}} \times EXP \left\{ -\frac{1}{2} (\mathbf{Y}_i - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}_0^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_0) \right\} \right)^{(1-d_i)}$$

Prior distributions

Hyperprior

$\theta \sim$ Uniform(0.1,0.9)

Priors

$D_i \sim$ Bernoulli(θ) (Observation i: 1,. . . ,N)

Se = Sp $\sim$ Beta(1,1)T(0.51,∞)

Prior distributions

Set $\mathbf{\Sigma}_j = \mathbf{V}_j \mathbf{R}_j \mathbf{V}_j^*$

For: $\mathbf{V}_j = \sigma_{k,j} I_3$ and $\mathbf{R}_j$ is a correlation matrix. [j:0,1; k:1,. . .,3]

Then: $\mathbf{C}_j = \begin{pmatrix} 1 & c_{j,12} & c_{j,13} \\ 0 & c_{j,22} & c_{j,23} \\ 0 & 0 & c_{j,33} \end{pmatrix}$ = Cholesky factor of $\mathbf{R}_j$.

$\sigma_{k,j} \sim$ Uniform(0,1000)

$c_{j,12} = \rho_{j,12} \sim$ Uniform(-1,1)

$c_{j,13} = \rho_{j,13} \sim$ Uniform(-1,1)

$c_{j,23} \sim \text{Uniform}\left(-\sqrt{1-\rho_{j,13}^2}, \sqrt{1-\rho_{j,13}^2}\right)$

$\rho_{j,23} = \rho_{j,12}\rho_{j,13} + c_{j,22}c_{j,23}$

* Wei, Y and Higgins, J.P.T (2013)

Prior distributions

$$AUC_{Index} = \Phi\left\{[(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})'(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})]^{\frac{1}{2}}\right\}$$

Reparameterize:

$$AUC_{Index} = \Phi\left\{\sqrt{\boldsymbol{\Delta}'\boldsymbol{\Delta}}\right\}$$

Where $\boldsymbol{\Delta} = \mathbf{L}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})$

For $\mathbf{L}$ = the Cholesky factor of $(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}$

Priors

$\boldsymbol{\Delta} \sim N_3(\kappa, \Psi)$

$\mu_{0k} \sim N(0, 10^6)$ (k: 1,...,3)

$\boldsymbol{\mu_1} = \boldsymbol{\Delta} L^{-1} + \boldsymbol{\mu_0}$

Bayesian latent-class mixture model

$\mathbf{Y}_{Index} = \hat{\mathbf{a}}'\mathbf{Y}_{Val}$ (Biomarker index observations)

Full data likelihood

$$L(\mu_0, \mu_1, \sigma_0, \sigma_1, \theta, Se, Sp | \mathbf{Y}_{Index}, \mathbf{T}_{Val}, \mathbf{D}_{Val})$$

$$= \prod_{i=1}^{N} \left(\theta Se^{t_{Val_i}} (1 - Se)^{(1-t_{Val_i})} \frac{1}{\sqrt{2\pi\sigma_1^2}} \times EXP\left\{ -\frac{(Y_{Index_i} - \mu_1)^2}{\sigma_1^2} \right\} \right)^{d_{Val_i}}$$

$$\times \left((1-\theta)(1-Sp)^{t_{Val_i}} Sp^{(1-t_{Val_i})} \frac{1}{\sqrt{2\pi\sigma_0^2}} \times EXP\left\{ -\frac{(Y_{Index_i} - \mu_0)^2}{\sigma_0^2} \right\} \right)^{(1-d_{Val_i})}$$

Prior distributions

Hyperprior

$\theta \sim$ Uniform(0.1,0.9)

Priors

$D_i \sim$ Bernoulli(θ) (Observation i: 1,. . . ,N)

Se = Sp $\sim$ Beta(1,1)T(0.51,∞)

$\sigma_j \sim$ Uniform(0,1000) [j:0,1]

Prior distributions

$$AUC_{Index} = \Phi\left\{\frac{(\mu_1-\mu_0)}{\sqrt{\sigma_0^2+\sigma_1^2}}\right\}$$

Reparameterize:

$$AUC_{Index} = \Phi\{\gamma\}$$
$$\text{Where } \gamma = \frac{(\mu_1-\mu_0)}{\sqrt{\sigma_0^2+\sigma_1^2}}$$

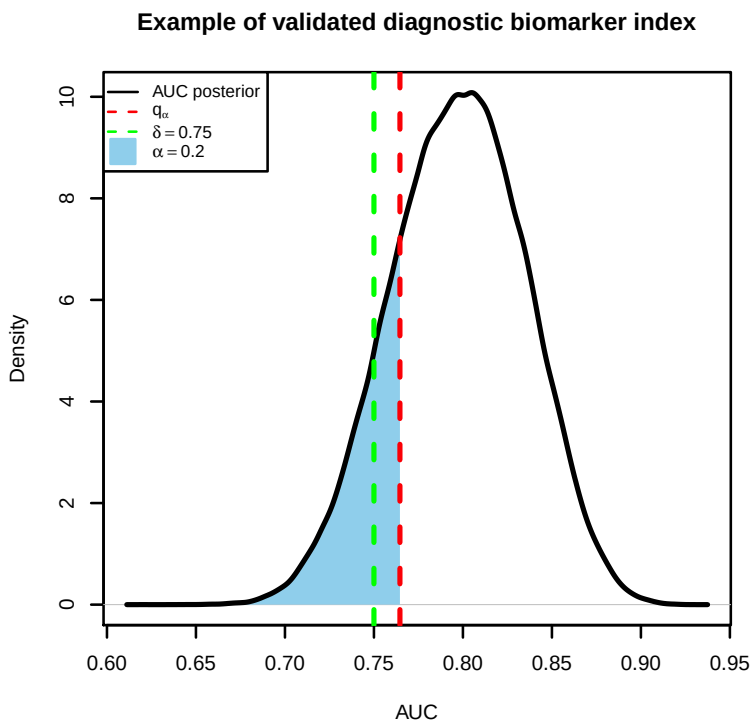
Priors

$$\gamma \sim N(\lambda, \tau^2)$$
$$\mu_0 \sim N(0, 10^6)$$
$$\mu_1 = \gamma \times \sqrt{\sigma_0^2 + \sigma_1^2} + \mu_0$$

Validation criterion

Based on Bayesian hypothesis testing paradigm:

$$H_0 : AUC \leq \delta$$
$$H_1 : AUC > \delta$$



Consider result significant when posterior probability of AUC exceeding δ is larger than $1 - \alpha$.

Incorporate pre-validation information

Development $\approx$ **Validation**

$$\phi^{-1}\{AUC_{Index}\} = \sqrt{\Delta'\Delta} \quad \approx \quad \gamma = \phi^{-1}\{AUC_{Index}\}$$

Take approximation to posterior distribution of $\sqrt{\Delta'\Delta}$ as prior distribution for γ

Incorporate pre-validation information

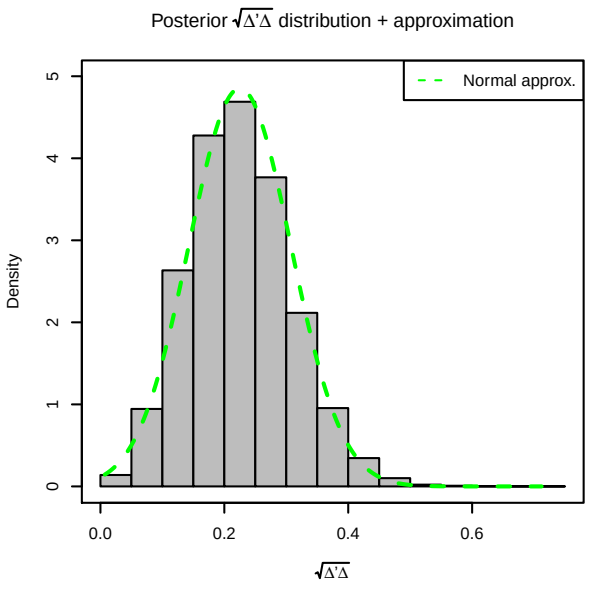
Development $\phi^{-1}\{AUC_{Index}\} = \sqrt{\Delta'\Delta}$ $\approx$ **Validation** $\gamma = \phi^{-1}\{AUC_{Index}\}$

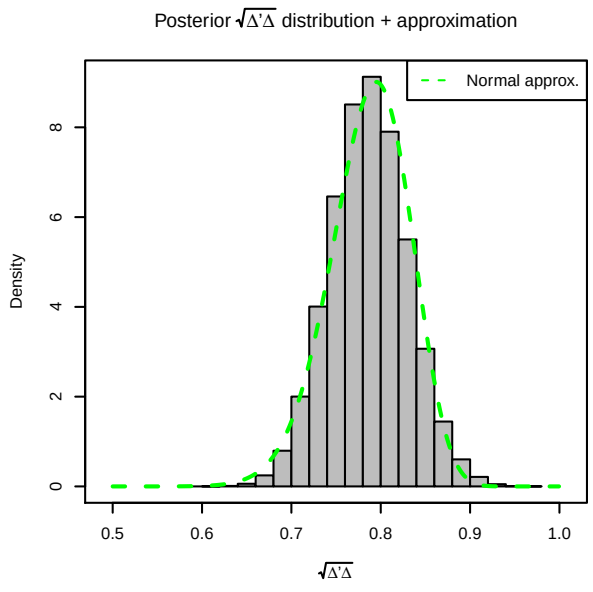
Take approximation to posterior distribution of $\sqrt{\Delta'\Delta}$ as prior distribution for γ

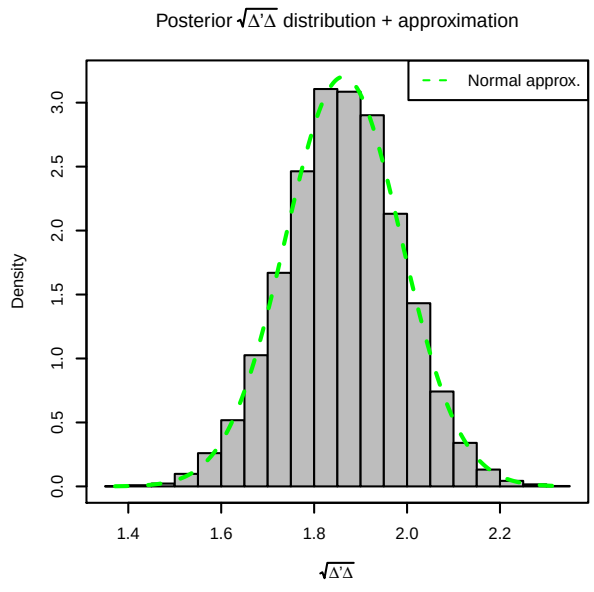
$\sqrt{\Delta'\Delta}$ posterior approximation

Prior: $\gamma \sim N(\lambda, \tau^2)$

Where $\lambda = \overline{x}_{\sqrt{\Delta'\Delta}_{1:M}}$, and $\tau^2 = s^2_{\sqrt{\Delta'\Delta}_{1:M}}$ (For M MCMC samples)







Outline

GOAL

Establish difference in power to validate AUC of biomarker index when ignoring vs incorporating pre-validation information

For 3 correlated biomarkers

$\theta = 0.5$

Se = Sp = 0.85

Mixture component parameters set such that:

AUC of biomarker 1 = 0.75
AUC of biomarker 2 = 0.75
AUC of biomarker 3 = 0.75

$AUC_{Index} = 0.78$

Development Stage

- $N_{Dev} = 400$
- $\hat{\mathbf{a}}'$ = posterior median of $\mathbf{a}'$

Validation Stage

- 200 data sets for $N_{Val} = 100, 400, 600$, and 800
- Power = proportion of simulations for which P(AUC > 0.75|data) > 0.80
- Prior AUC information:
  - Ignoring AUC information (Red)
    - Prior γ : $N(0, 1)$
  - Incorporating AUC information (Blue)
    - Prior γ : Discounted normal approx. to posterior predictive distribution of $\sqrt{\Delta' \Delta}$

Development Stage

- $N_{Dev} = 400$
- $\hat{\mathbf{a}}'$ = posterior median of $\mathbf{a}'$

Validation Stage

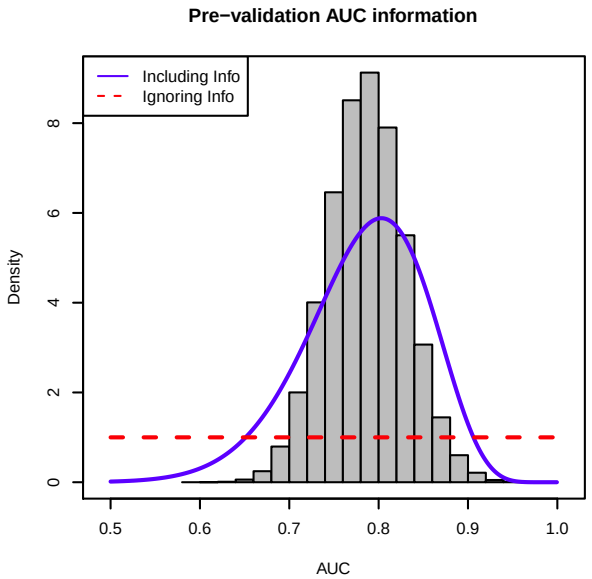
- 200 data sets for $N_{Val} = 100$, 400, 600, and 800
- Power = proportion of simulations for which P(AUC > 0.75|data) > 0.80
- Prior AUC information:
  - Ignoring AUC information (Red)
    - Prior γ : $N(0, 1)$
  - Incorporating AUC information (Blue)
    - Prior γ : Discounted normal approx. to posterior predictive distribution of $\sqrt{\Delta' \Delta}$

Development Stage

- $N_{Dev} = 400$
- $\hat{\mathbf{a}}'$ = posterior median of **a**'

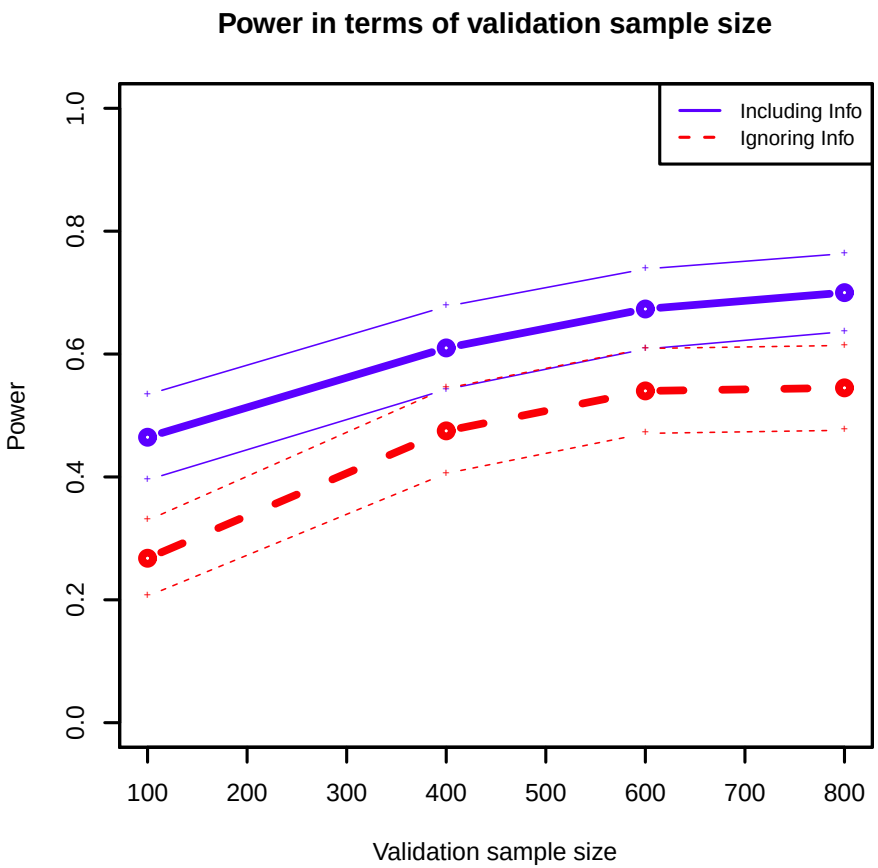
Validation Stage

- 200 data sets for $N_{Val} = 100$, 400, 600, and 800
- Power = proportion of simulations for which P(AUC > 0.75|data) > 0.80
- Prior AUC information:



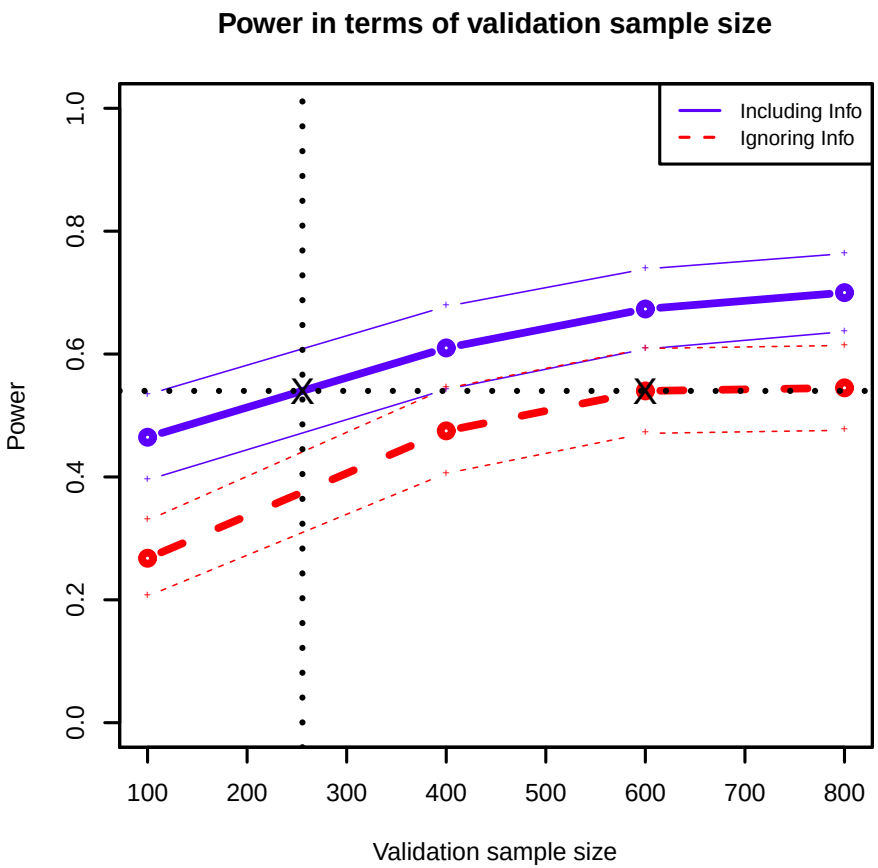
- Ignoring AUC information (Red)
  - Prior γ : $N(0, 1)$
- Incorporating AUC information (Blue)
  - Prior γ : Discounted normal approx. to posterior predictive distribution of $\sqrt{\Delta'\Delta}$

Simulated Power



- Increasing validation sample size increases power
- Incorporating pre-validation information significantly increases power
- Reduction of about $\frac{1}{2}$ of sample size to maintain power

Simulated Power



- Increasing validation sample size increases power
- Incorporating pre-validation information significantly increases power
- Reduction of about $\frac{1}{2}$ of sample size to maintain power

Outline

Conclusions

- Bayesian framework:
  - Allows including pre-validation information into validation stage
    - By approximating posterior AUC information as prior
    - Also other pre-validation information possible
- Simulation study
  - Power to reach validation is significantly increased
  - Sample size reduction for equal power

Further considerations

- Other validation criteria
- Robustness to miss-estimated linear combination coefficients
- Extend to incorporate non-normally distributed biomarkers
- Evaluate impact of conditional independence assumption

References

- Su, J.Q., Liu, J.S.: Linear combinations of multiple diagnostic markers. Journal of the American Statistical Association. **88**, 1350–1355 (1993)
- Wei, Y, Higgins, P.T.: Bayesian multivariate meta-analysis with multiple outcomes. Statistics in Medicine (2013) doi: 10.1002/sim.5745

Thank you for your attention !