Estimation of Optimally-Combined-Biomarker Accuracy in the Absence of a Gold-Standard Reference Test

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

Problem setting
- Accuracy definition
- Optimal combination of biomarkers
- Absence of gold-standard reference

Bayesian latent-class mixture model
- ”Naive” prior definition
- Controlled prior definition

Simulation study
- Data
- Results

Conclusions

Outline

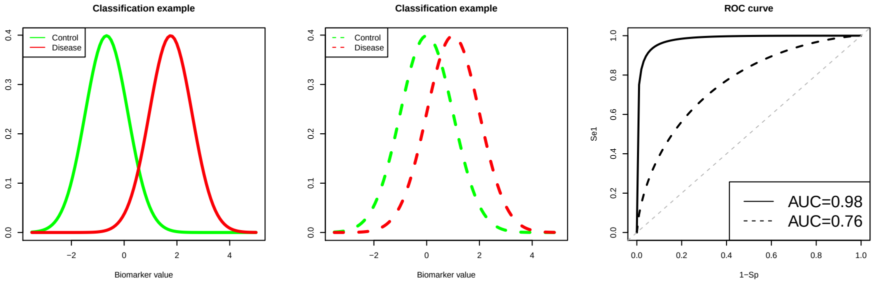
Problem setting

Establish accuracy of a combination of biomarkers in the absence of a gold-standard reference test

- Area under the Receiver Operating Characteristics (ROC) curve (AUC) as measure of accuracy
- Choose combination of biomarkers that maximizes AUC
- Imperfect reference test leads to biased estimates of accuracy

=> To this end a Bayesian latent-class mixture model 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)
  - $\mathbf{Y}|_{D=0} \sim N_K(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$
  - $\mathbf{Y}|_{D=1} \sim N_K(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
- Se: Unknown sensitivity of the reference test (T)
- Sp: Unknown specificity of the reference test (T)
- θ: Unknown true prevalence of disease in the data set
- Reference test is imperfect
  - Conditionally on true disease status, misclassification independent of biomarker value
  - Ignoring will UNDERESTIMATE performance of biomarker

ROC parameters optimal combination of biomarkers

According to Siu and Liu (1993) the linear combination maximizing AUC is 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:

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

This is all under the assumption of a gold standard reference test. We propose to extend this to the imperfect reference test case.

ROC parameters optimal combination of biomarkers

According to Siu and Liu (1993) the linear combination maximizing AUC is 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:

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

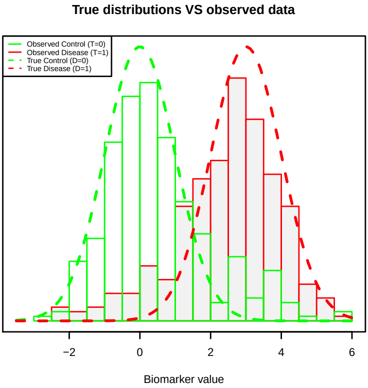
Area Under the ROC Curve:

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

This is all under the assumption of a gold standard reference test. We propose to extend this to the imperfect reference test case.

Underlying versus observed data

Ignoring misclassification in imperfect reference test will lead to bias of estimated accuracy:



- In example: conditionally independent misclassification
- Misclassification in reference test causes skewed observed distributions
- Goal: retrieve accuracy of true underlying biomarker by observed data

Outline

- Problem setting
  - Accuracy definition
  - Optimal combination of biomarkers
  - Absence of gold-standard reference
- Bayesian latent-class mixture model
  - ”Naive” prior definition
  - Controlled prior definition
- Simulation study
  - Data
  - Results
- Conclusions

Full data likelihood

$$
\begin{aligned}
&L(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \theta, Se, Sp | \mathbf{Y}, \mathbf{T}, \mathbf{D}) \\
&= \prod_{i=1}^{N} \left(\theta Se^{t_i} (1 - Se)^{(1-t_i)} \frac{1}{\sqrt{2\pi |\boldsymbol{\Sigma}_1|}} \times EXP \left\{ -\frac{1}{2} (\mathbf{Y}_i - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_1) \right\} \right)^{d_i} \\
&\times \left((1-\theta)(1 - Sp)^{t_i} Sp^{(1-t_i)} \frac{1}{\sqrt{2\pi |\boldsymbol{\Sigma}_0|}} \times EXP \left\{ -\frac{1}{2} (\mathbf{Y}_i - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}_0^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_0) \right\} \right)^{(1-d_i)}
\end{aligned}
$$

"Naive" prior definition

Hyperprior

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

Priors

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

$\mu_{kj} \sim$ N(0,10^6) (Disease indicator j: 0, 1; Biomarker k: 1,. . .,K)

$\boldsymbol{\Sigma}_j^{-1} \sim$ Wish($\boldsymbol{S}$,K) (Disease indicator j: 0, 1)

with **S** = VarCov-matrix of observed control group

Se = Sp $\sim$ Beta(1,1)T(0.51,∞) [Non-informative]

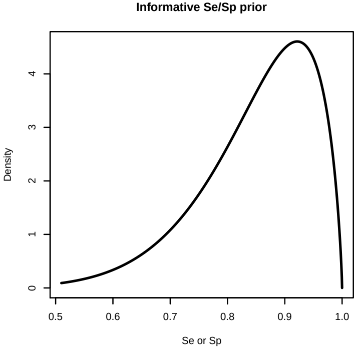
OR Se = Sp $\sim$ Beta(10,1.764706)T(0.51,∞) [Informative]

Se/Sp Beta(10,1.764706) Prior

Mean = 0.85

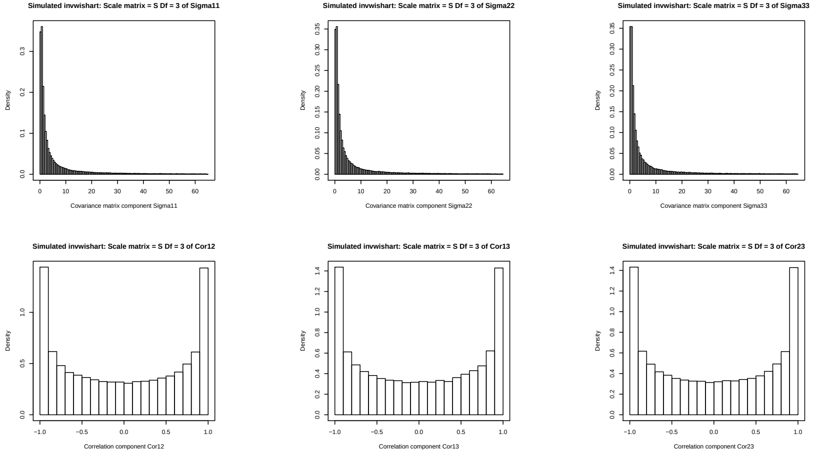
Var = 0.009988479

Equal-tail 95%-probability interval: 0.6078 - 0.9834



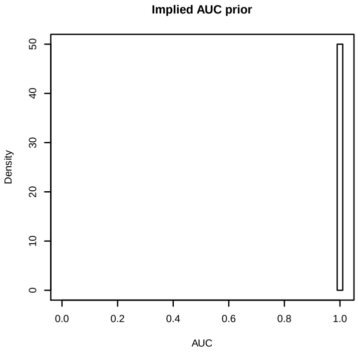
Implied priors

Variances and correlations



Implied priors

AUC



- Prior specification is used commonly (e.g. O'Malley and Zou (2006))
- Uninformative mixture component priors lead to prior point mass distribution centred at 1 for AUC
- Extremely informative prior for component of interest!

Controlled prior definition ($\boldsymbol{\Sigma}$)

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

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

Then: $\mathbf{C}_j$ = Cholesky factor of $\mathbf{R}_j$.

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

Say K=3 then:

$$C_{j,12} = \rho_{j,12} \sim \text{Uniform(-1,1)}$$
$$C_{j,13} = \rho_{j,13} \sim \text{Uniform(-1,1)}$$
$$C_{j,23} \sim \text{Uniform}\left(-\sqrt{1-\rho_{j,13}^2}, \sqrt{1-\rho_{j,13}^2}\right)$$
$$\rho_{j,23} = \rho_{j,12}\rho_{j,13} + C_{j,22}C_{j,23}$$

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

Controlled prior definition (AUC)

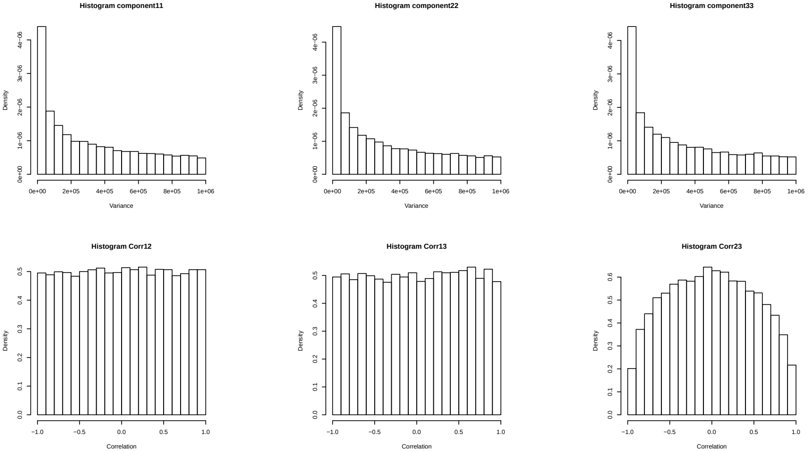
Set $\mathbf{\Delta} = \mathbf{L}(\boldsymbol{\mu}_\mathbf{1} - \boldsymbol{\mu}_\mathbf{0})$

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

$$\mathbf{\Delta} \sim N_K(\boldsymbol{\kappa}, \Psi)$$
$$\mu_{0k} \sim N(0, 10^6) \ \ \text{(k: 1,...,K)}$$
$$\boldsymbol{\mu}_1 = \mathbf{\Delta} L^{-1} + \boldsymbol{\mu}_\mathbf{0}$$

Implied priors

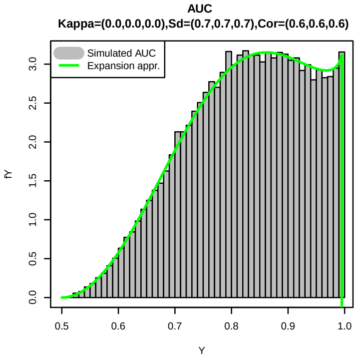
Variances and correlations



Implied priors

AUC

For $\boldsymbol{\kappa} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, $\sigma_i = 0.7$ and $\rho_{ij} = 0.6$ [i,j: 1,. . .,K]



- Less informative prior distribution for AUC
- Prior on $\boldsymbol{\Delta}$ gives control over informativeness AUC prior

Outline

400 datasets for 3 independent biomarkers

N = 100, 400 or 600
$\theta = 0.5$
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

$AUC_{OptimalCombination} = 0.88$

AUC Results (Average of median posterior AUC)

True AUC = 0.8786

Prior Formulation	Se/Sp Prior	Sample Size		
		N=100	N=400	N=600
GS	/	0.7710 (0.0361)	0.7661 (0.0210)	0.7614 (0.0157)

- Gold Standard model fit leads to severe underestimation
- Naive AUC prior specification causes slight overestimation
  - Increased sample size reduces overestimation and decreases standard errors
  - Informative Se/Sp prior also reduces this bias, but seems to increase standard errors
- Controlled AUC prior reduces overestimation compared to Naive-prior case
  - Increased sample size decreases standard errors
  - Informative Se/Sp prior no substantial effect

AUC Results (Average of median posterior AUC)

True AUC = 0.8786

Prior Formulation	Se/Sp Prior	Sample Size		
		N=100	N=400	N=600
GS	/	0.7710 (0.0361)	0.7661 (0.0210)	0.7614 (0.0157)
Naive	**Non-Inf**	0.9241 (0.0279)	0.8890 (0.0279)	0.8836 (0.0262)
Naive	**Inf**	0.9068 (0.0344)	0.8827 (0.0286)	0.8785 (0.0263)

- Gold Standard model fit leads to severe underestimation
- Naive AUC prior specification causes slight overestimation
  - Increased sample size reduces overestimation and decreases standard errors
  - Informative Se/Sp prior also reduces this bias, but seems to increase standard errors
- Controlled AUC prior reduces overestimation compared to Naive-prior case
  - Increased sample size decreases standard errors
  - Informative Se/Sp prior no substantial effect

AUC Results (Average of median posterior AUC)

True AUC = 0.8786

Prior Formulation	Se/Sp Prior	Sample Size		
		N=100	N=400	N=600
GS	/	0.7710 (0.0361)	0.7661 (0.0210)	0.7614 (0.0157)
Naive	**Non-Inf**	0.9241 (0.0279)	0.8890 (0.0279)	0.8836 (0.0262)
Naive	**Inf**	0.9068 (0.0344)	0.8827 (0.0286)	0.8785 (0.0263)
Controlled	**Non-Inf**	0.8907 (0.0347)	0.8803 (0.0290)	0.8773 (0.0271)
Controlled	**Inf**	0.8728 (0.0388)	0.8741 (0.0292)	0.8722 (0.0269)

- Gold Standard model fit leads to severe underestimation
- Naive AUC prior specification causes slight overestimation
  - Increased sample size reduces overestimation and decreases standard errors
  - Informative Se/Sp prior also reduces this bias, but seems to increase standard errors
- Controlled AUC prior reduces overestimation compared to Naive-prior case
  - Increased sample size decreases standard errors
  - Informative Se/Sp prior no substantial effect

Outline

Conclusions

- Bayesian latent-class mixture model:
  - Takes unknown true disease status into account
  - Incorporates information from reference test while acknowledges imperfectness
    - Provides estimates of accuracy of the reference test
- Simulation study
  - Model is able to retrieve true AUC
- Careful prior specification
  - Complex function of uninformative prior distributions => informative prior => biased estimates
  - Controlled prior specification is proposed

Further considerations

- Sensitivity to misspecified Se/Sp prior distribution
- Extend to incorporate non-normally distributed biomarkers
- Evaluate impact of conditional independence assumption

References

- O'Malley, A.J., Zou, K.H.: Bayesian multivariate hierarchical transformation models for ROC analysis. Statistical Medicine. **25**, 459–479 (2006)
- 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 !