Obtaining a candidate set of covariates for genetic risk scores under data protection constraints in consortia

Daniela Zöller[1] Harald Binder[1]
Marcel Schniedermann[2] Dennis Kadioglu[1]

[1]Institute of Medical Biostatistics, Epidemiology and Informatics,
[2]Johannes Gutenberg University Mainz

November 11th, 2016

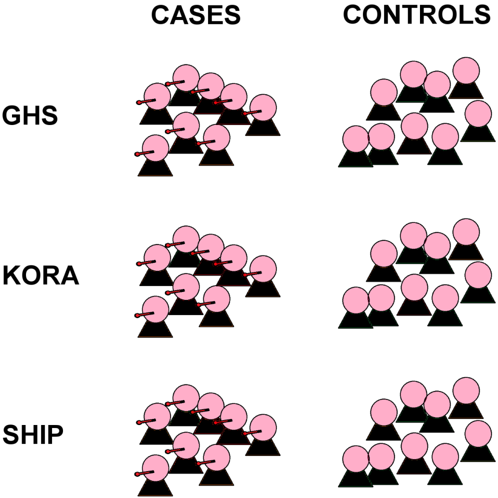
Genome-wide association study with focus on SNPs

Are there SNPs associated with disease status?

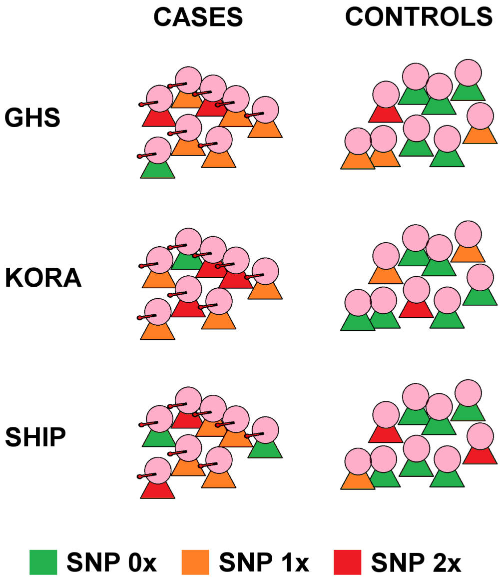
Can we obtain a genetic risk score?

$$E[y] = \mathbf{X}\underline{\beta}$$

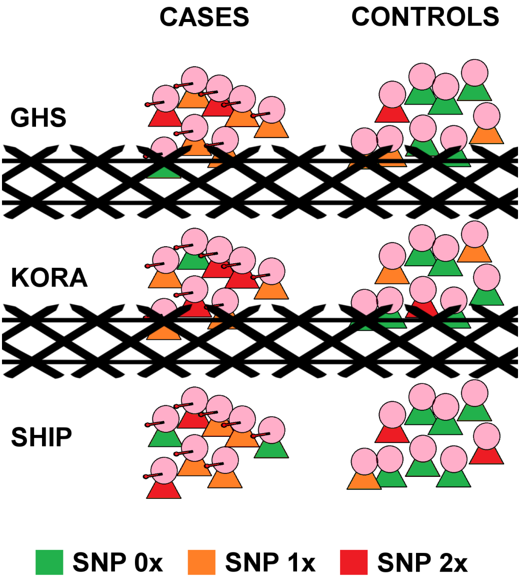
Genome-wide association study with focus on SNPs



Genome-wide association study with focus on SNPs



Genome-wide association study with focus on SNPs



Goal: Multivariable genetic risk score

- Pool data from different cohorts
- Problem: Data protection constraints
  - Individual data not available for analysis
  - Aggregated data available
- Problem: High number of potential (correlated) SNPs

Our proposal: Distributed Boosting

Multivariable regularized stagewise regression model
incorporating automatic variable selection
using univariable effect estimates
and covariances of data

Goal: Multivariable genetic risk score

- Pool data from different cohorts
- Problem: Data protection constraints
  - Individual data not available for analysis
  - Aggregated data available
- Problem: High number of potential (correlated) SNPs

Our proposal: Distributed Boosting

Multivariable regularized stagewise regression model
incorporating automatic variable selection
using univariable effect estimates
and covariances of data

Goal: Multivariable genetic risk score

- Pool data from different cohorts
- Problem: Data protection constraints
  - Individual data not available for analysis
  - Aggregated data available
- Problem: High number of potential (correlated) SNPs

Our proposal: Distributed Boosting

Multivariable regularized stagewise regression model
incorporating automatic variable selection
using univariable effect estimates
and covariances of data

Goal: Multivariable genetic risk score

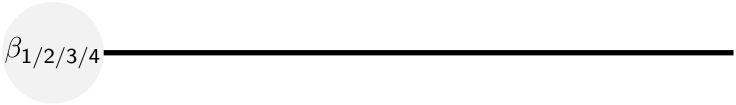
- Pool data from different cohorts
- Problem: Data protection constraints
  - Individual data not available for analysis
  - Aggregated data available
- Problem: High number of potential (correlated) SNPs

Our proposal: Distributed Boosting

Multivariable regularized stagewise regression model
incorporating automatic variable selection
using univariable effect estimates
and covariances of data

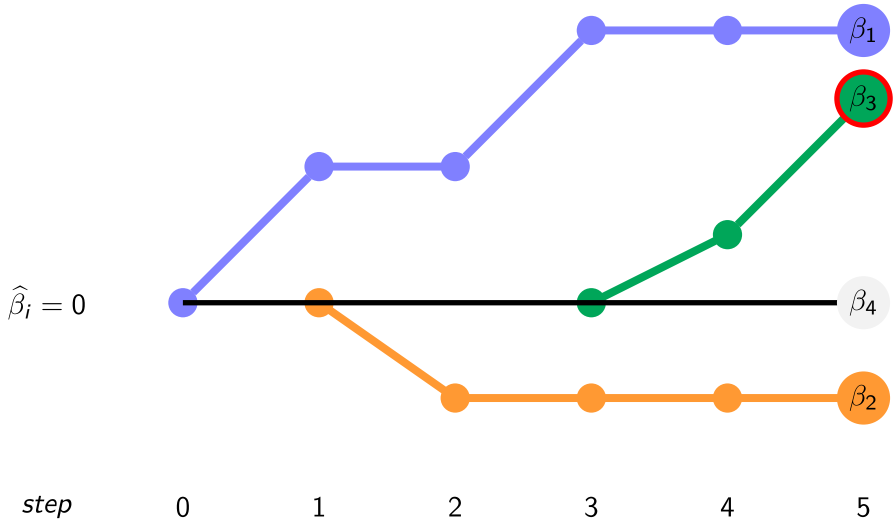
Boosting

$\widehat{\beta}_i = 0$



step 0 1 2 3 4 5

Boosting



Needed data for update in step (m+1) in a linear regression model

Score function for covariate j:

$$\widetilde{S}_j^{(m+1)} = x_j' r_{(j)}^{(m)}$$

$$= \sum_{i=1}^{n} x_{ij} \left(y_i - \widehat{y}_i^{(m)} \right)$$

$$= \sum_{i=1}^{n} x_{ij} y_i - \sum_{k:|\widehat{\beta}_k^{(m)}|>0} \widehat{\beta}_k^{(m)} \sum_{i=1}^{n} x_{ij} x_{ik}$$

$$\underset{\text{Univariable estimate}}{\uparrow} \qquad \underset{\text{Covariance of covariates}}{\uparrow}$$

Needed data for update in step (m+1) in a linear regression model

Score function for covariate j:

$$
\begin{aligned}
\widetilde{S}_j^{(m+1)} &= x_j' r_{(j)}^{(m)} \\
&= \sum_{i=1}^{n} x_{ij} \left(y_i - \widehat{y}_i^{(m)} \right) \\
&= \sum_{i=1}^{n} x_{ij} y_i - \sum_{k:|\widehat{\beta}_k^{(m)}|>0} \widehat{\beta}_k^{(m)} \sum_{i=1}^{n} x_{ij} x_{ik}
\end{aligned}
$$

$\uparrow$ Univariable estimate

$\uparrow$ Covariance of covariates

Needed data for update in step (m+1) in a linear regression model

Score function for covariate j:

$$
\begin{aligned}
\widetilde{S}_j^{(m+1)} &= x_j' r_{(j)}^{(m)} \\
&= \sum_{i=1}^{n} x_{ij} \left(y_i - \widehat{y}_i^{(m)} \right) \\
&= \underbrace{\sum_{i=1}^{n} x_{ij} y_i}_{\uparrow\ \text{Univariable estimate}} - \sum_{k:|\widehat{\beta}_k^{(m)}|>0} \widehat{\beta}_k^{(m)} \underbrace{\sum_{i=1}^{n} x_{ij} x_{ik}}_{\uparrow\ \text{Covariance of covariates}}
\end{aligned}
$$

Needed data for update in step (m+1) in a linear regression model

Score function for covariate j:

$$
\begin{aligned}
\widetilde{S}_j^{(m+1)} &= x_j' r_{(j)}^{(m)} \\
&= \sum_{i=1}^{n} x_{ij} \left(y_i - \widehat{y}_i^{(m)} \right) \\
&= \underbrace{\sum_{i=1}^{n} x_{ij} y_i}_{\substack{\uparrow \\ \text{Univariable} \\ \text{estimate}}} - \sum_{k:|\widehat{\beta}_k^{(m)}|>0} \widehat{\beta}_k^{(m)} \underbrace{\sum_{i=1}^{n} x_{ij} x_{ik}}_{\substack{\uparrow \\ \text{Covariance} \\ \text{of covariates}}}
\end{aligned}
$$

Assumptions

- Multivariate linear regression model
  - Response and covariates standardized
- Available data
  - Univariable estimates obtained from linear regression models
  - Covariance of already selected covariates with all covariates
- Several data providers
  - Pool univariable estimates and covariances using weighted mean

Assumptions

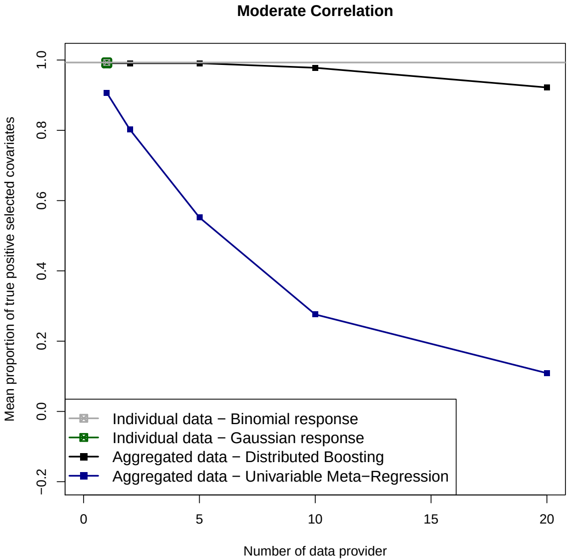
- Multivariate linear regression model
  - Response and covariates standardized
- Available data
  - Univariable estimates obtained from linear regression models
  - Covariance of already selected covariates with all covariates
- Several data providers
  - Pool univariable estimates and covariances using weighted mean

Assumptions

- Multivariate linear regression model
  - Response and covariates standardized
- Available data
  - Univariable estimates obtained from linear regression models
  - Covariance of already selected covariates with all covariates
- Several data providers
  - Pool univariable estimates and covariances using weighted mean

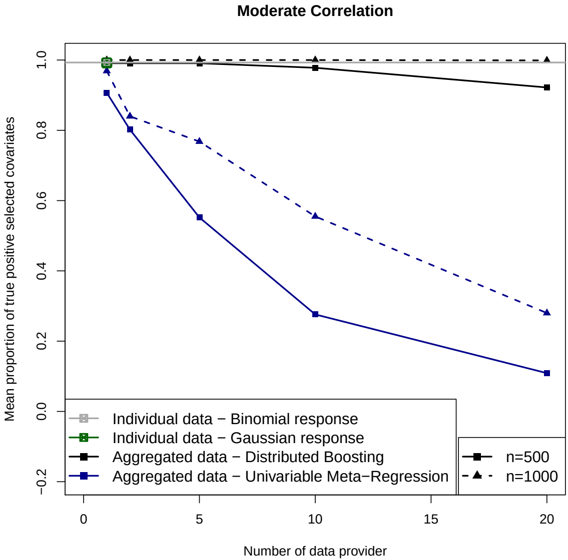
Mean proportion of true positive selected covariates

10 covariates with strong effect out of a total of 2500 covariates



Mean proportion of true positive selected covariates

10 covariates with strong effect out of a total of 2500 covariates



Privacy preserving approach

- Include trusted third party
  - Virtual or real
  - Set between analysis server and data provider
- Additional disassociation
  - Permute covariance matrix across data provider
  - Solemnly transferring the pooled data

Privacy preserving approach

- Include trusted third party
  - Virtual or real
  - Set between analysis server and data provider
- Additional disassociation
  - Permute covariance matrix across data provider
  - Solemnly transferring the pooled data

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...

[1]Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$ $<$	$S_{1,2}$ $>$	$S_{1,3}$ $>$	$S_{1,4}$	...

[1] Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...

[1]Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...
Step 2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	...

[1] Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...
Step 2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	...

[1] Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...
Step 2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	...

[1] Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...
Step 2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	...

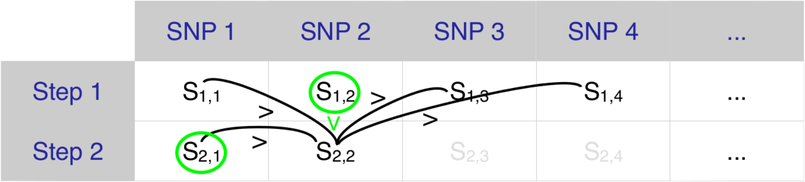
[1] Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...
Step 2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	...

[1] Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]



[1] Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...
Step 2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	...
Step 3	$S_{3,1}$	$S_{3,2}$	$S_{3,3}$	$S_{3,4}$	...

[1] Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...
Step 2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	...
Step 3	$S_{3,1}$	$S_{3,2}$	$S_{3,3}$	$S_{3,4}$	...

[1] Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...
Step 2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	...
Step 3	$S_{3,1}$	$S_{3,2}$	$S_{3,3}$	$S_{3,4}$	...

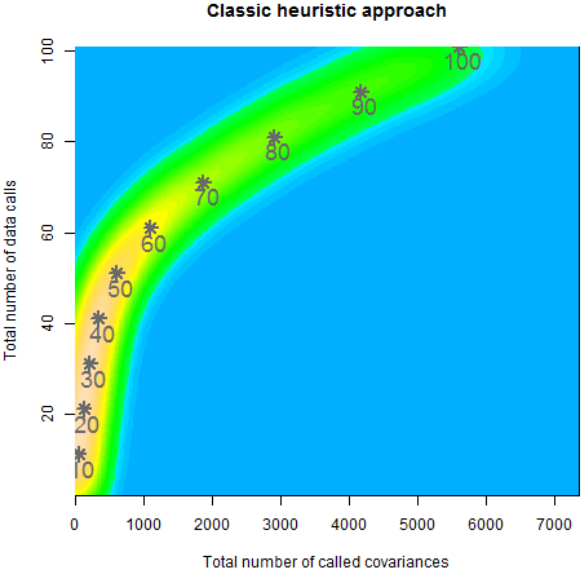
[1]Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

A heuristic for evaluating only some covariates[1]

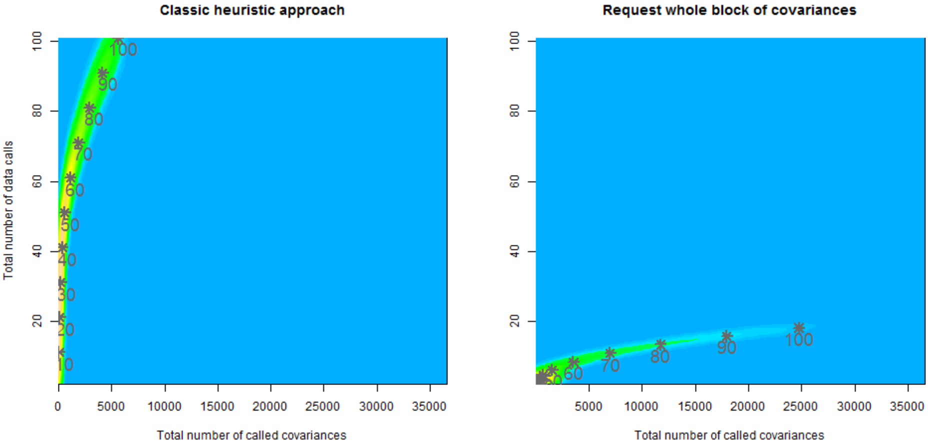
	SNP 1	SNP 2	SNP 3	SNP 4	...
Step 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	...
Step 2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	...
Step 3	$S_{3,1}$	$S_{3,2}$	$S_{3,3}$	$S_{3,4}$	...
...	$S_{4,1}$	$S_{4,2}$	$S_{4,3}$	$S_{4,4}$	...

[1] Binder et al. 2012: Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures; Statistics in Medicine 32 (10): 1778-1791.

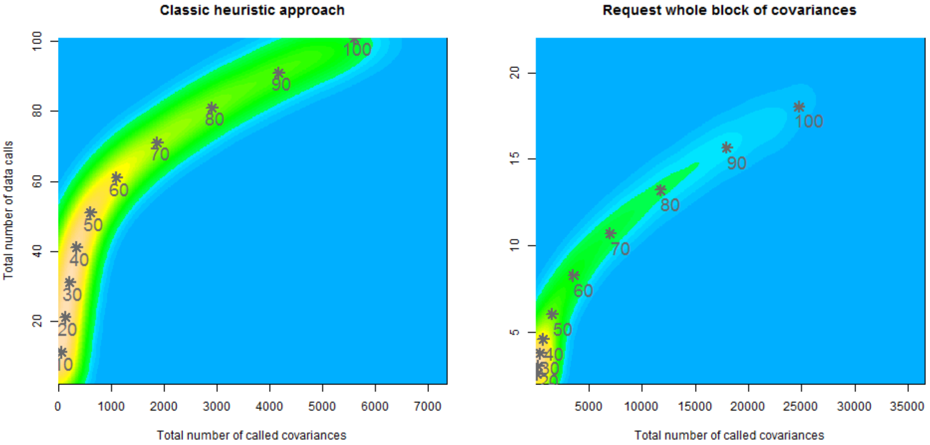
Amount of needed calls and data



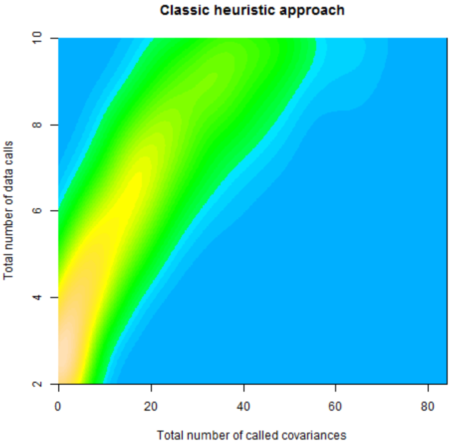
Amount of needed calls and data

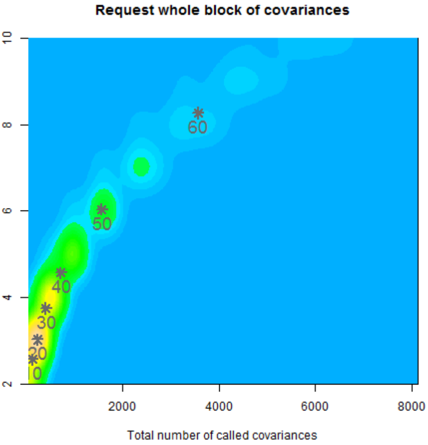


Amount of needed calls and data



Amount of needed calls and data





Conclusions

- Framework for high-dimensional multivariable analysis in consortia under data protection constraints
  - Result: Adequate genetic risk score
  - Only aggregated data needed
- Number and amount of data calls can be reduced
  - Release of data can be handled manually
- Automation in progress
  - Can be handled without much manpower
  - New patients can be added easily
- Exploratory mode possible

Conclusions

- Framework for high-dimensional multivariable analysis in consortia under data protection constraints
  - Result: Adequate genetic risk score
  - Only aggregated data needed
- Number and amount of data calls can be reduced
  - Release of data can be handled manually
- Automation in progress
  - Can be handled without much manpower
  - New patients can be added easily
- Exploratory mode possible

Conclusions

- Framework for high-dimensional multivariable analysis in consortia under data protection constraints
  - Result: Adequate genetic risk score
  - Only aggregated data needed
- Number and amount of data calls can be reduced
  - Release of data can be handled manually
- Automation in progress
  - Can be handled without much manpower
  - New patients can be added easily
- Exploratory mode possible

Conclusions

- Framework for high-dimensional multivariable analysis in consortia under data protection constraints
  - Result: Adequate genetic risk score
  - Only aggregated data needed
- Number and amount of data calls can be reduced
  - Release of data can be handled manually
- Automation in progress
  - Can be handled without much manpower
  - New patients can be added easily
- Exploratory mode possible

Thanks to...

Harald Binder, IMBEI Mainz
Dennis Kadioglu, IMBEI Mainz

Marcel Schniedermann, Uni Mainz

Contact
Daniela Zöller

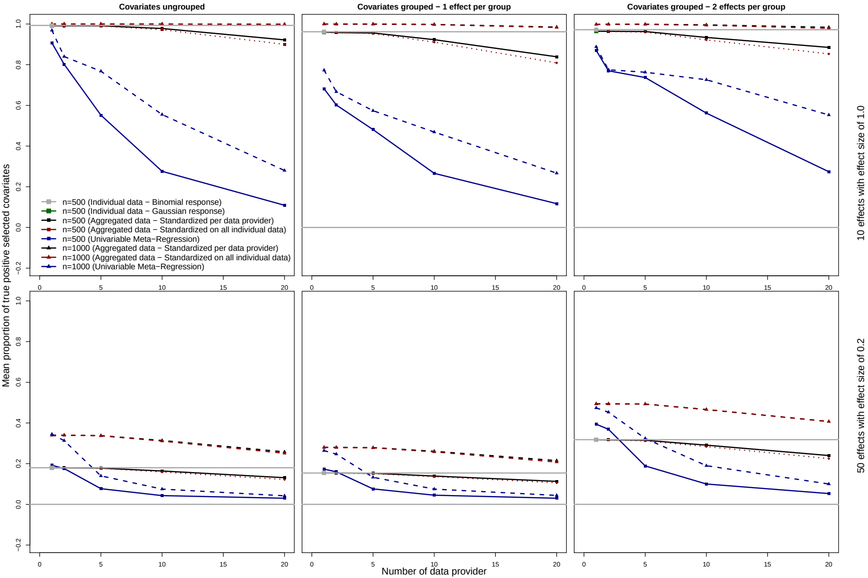
Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI)
University Medical Center
Johannes Gutenberg University Mainz, Germany

daniela.zoeller@uni-mainz.de

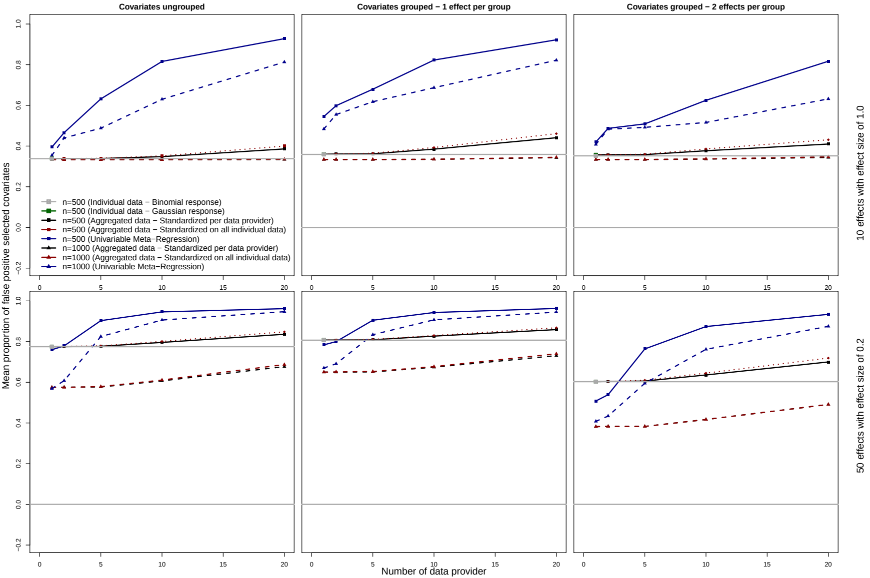
Simulation study

- Sample size: $n = 500/1000$ distributed on varying number of cohorts
- Number of covariates: $p = 2500$
- Correlation scenarios
  - Moderate correlation overall
    P(Same value for neighboring covariates)=0.5
  - Moderate correlation + groups of 5 with high correlation
    P(Same value for neighboring covariates | Diff. group) =0.5
    P(Same value for neighboring covariates | Same group) = 0.75
- Effect scenarios: Limited number of covariates with effect
  - 10 covariates with strong effect (effect size = 1)
  - 50 covariates with weak effect (effect size = 0.2)
- Number of covariates with effect per correlated group: 1 or 2
- Number of independent simulations: $M = 1000$

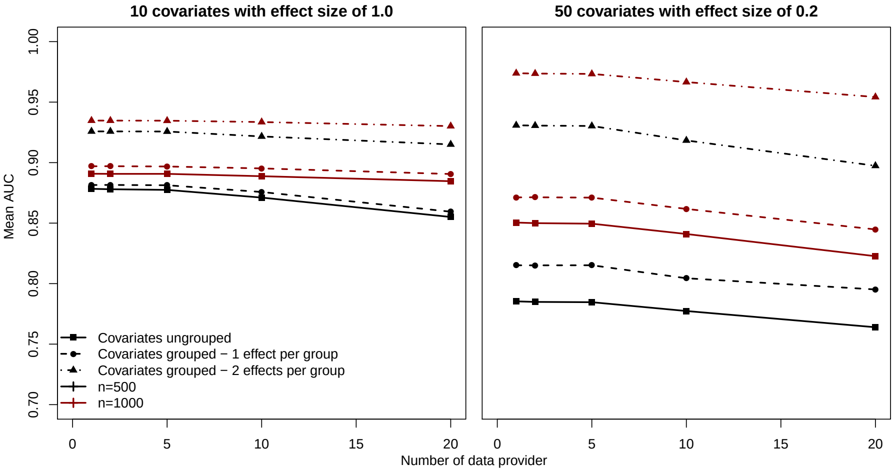
Mean proportion of true positive selected covariates



Mean proportion of false positive selected covariates



Mean AUC under ROC



Block-heuristic approach with varying buffer sizes