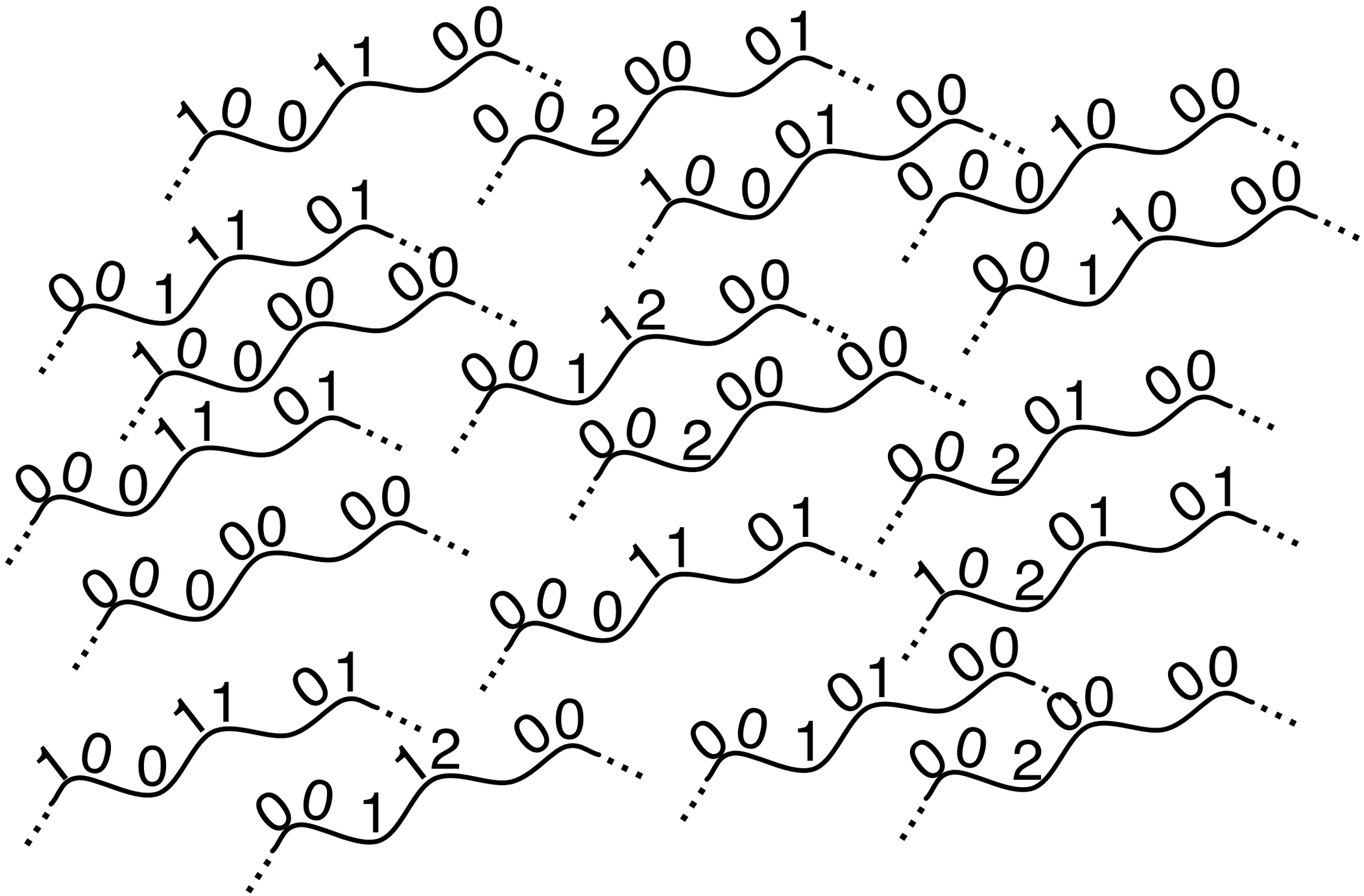
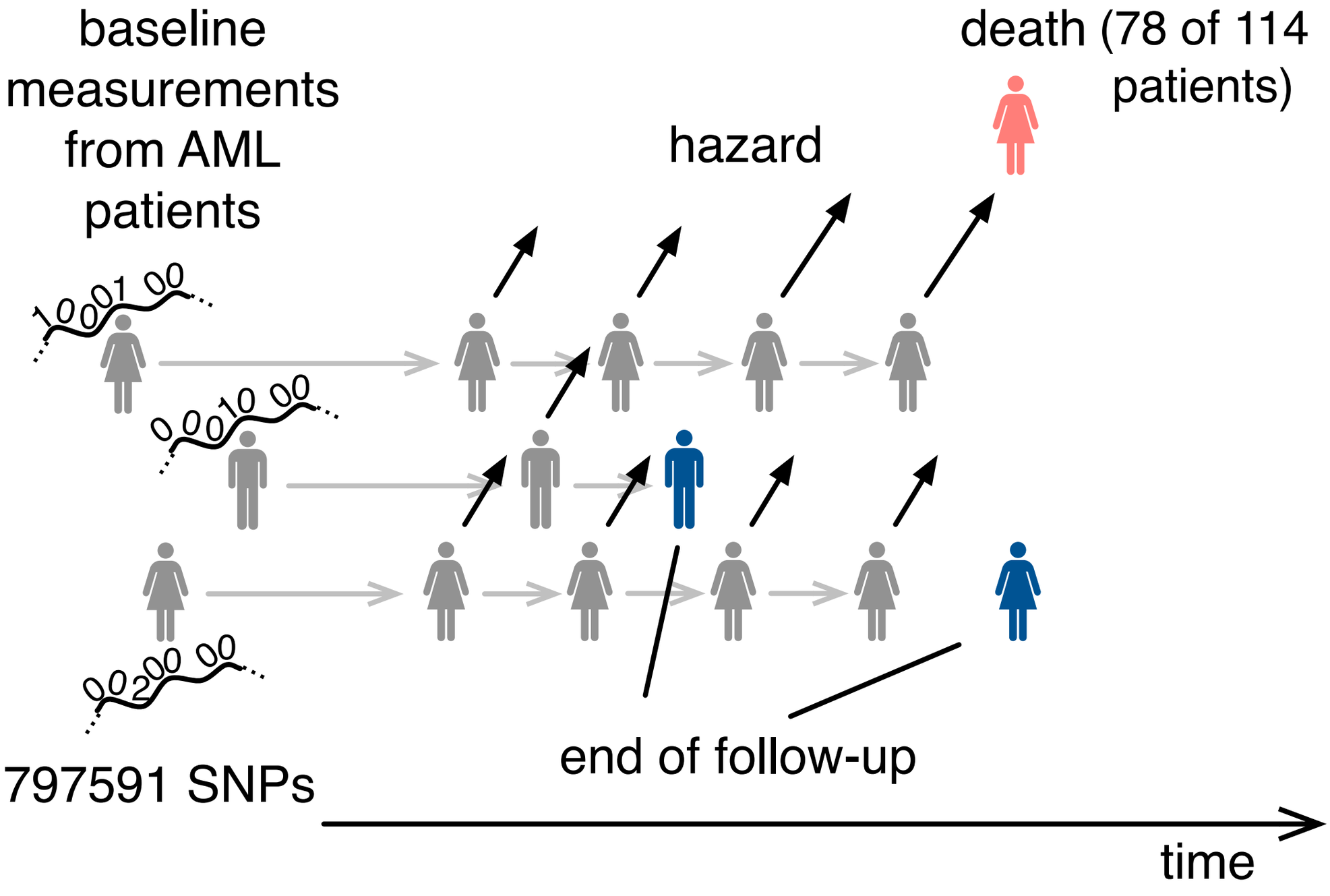
UNIVERSITÄTSmedizin.
MAINZ
Institut für Medizinische Biometrie,
Epidemiologie und Informatik

Single nucleotide polymorphisms (SNPs)



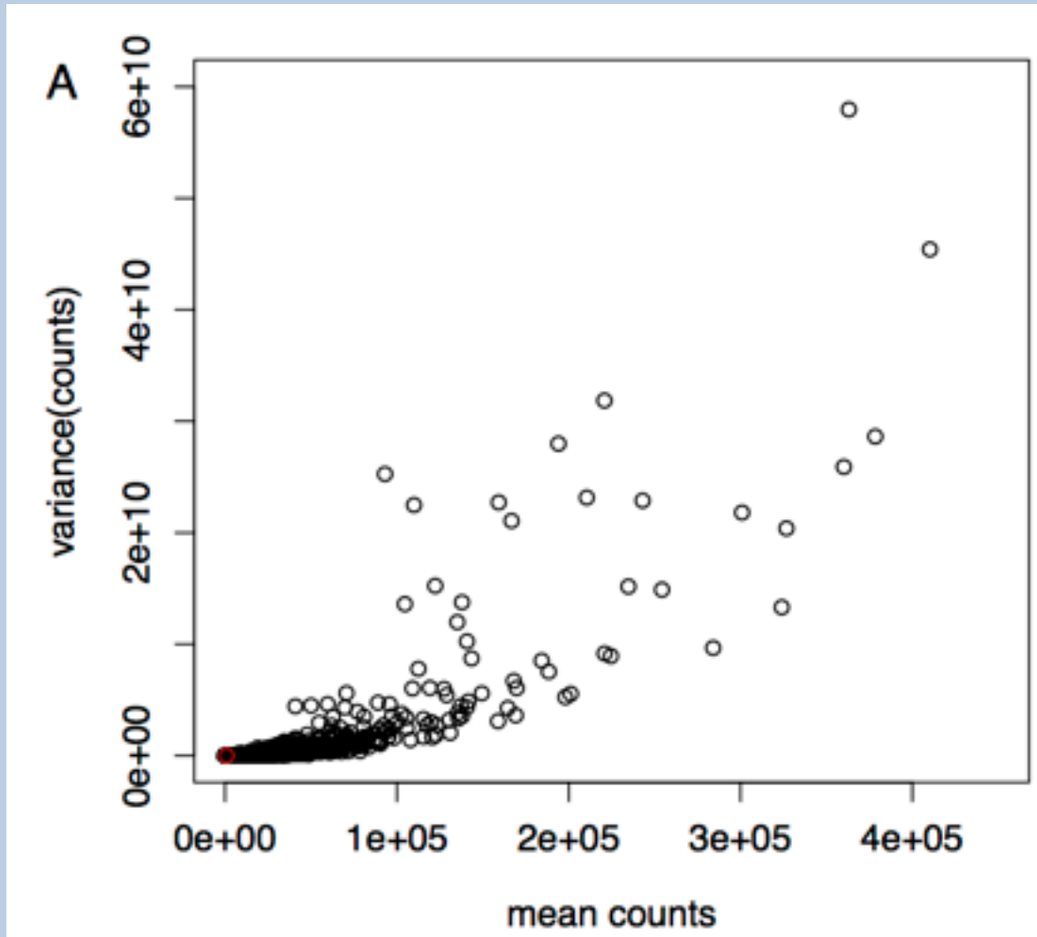
Institut für Medizinische Biometrie,
Epidemiologie und Informatik

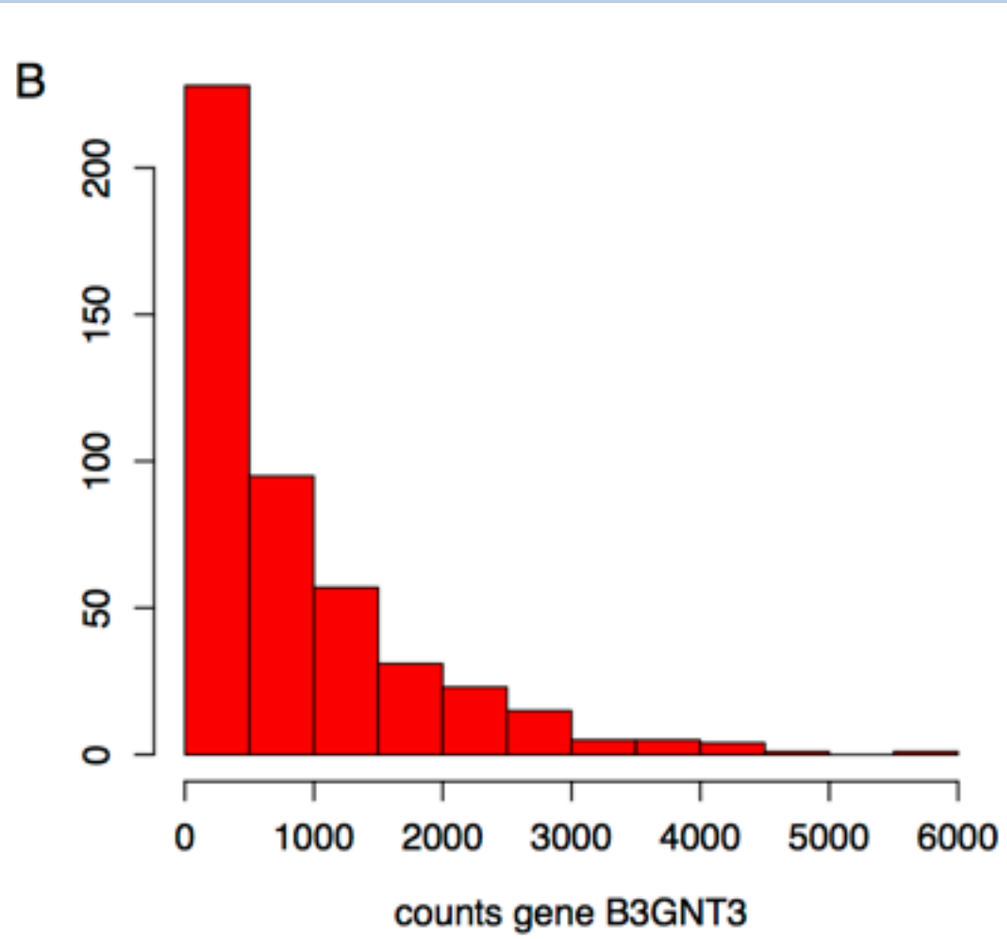
Prospective clinical cohort



RNA-Seq data

- 502 kidney renal clear cell carcinoma patients (from TCGA) with 160 deaths
- RNA-Seq measurements for 20,532 genes





JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Medizinische Biometrie,
Epidemiologie und Informatik

Risk prediction models

individual
risk

↗ $= \beta_1 \cdot 2 + \beta_2 \cdot 2 + \ldots + \beta_{797591} \cdot 0$

↗ $= \beta_1 \cdot 0 + \beta_2 \cdot 1 + \ldots + \beta_{797591} \cdot 1$

↗ $= \beta_1 \cdot 0 + \beta_2 \cdot 1 + \ldots + \beta_{797591} \cdot 0$

↗ $= \beta_1 \cdot 1 + \beta_2 \cdot 2 + \ldots + \beta_{797591} \cdot 2$

↗ $= \beta_1 \cdot 1 + \beta_2 \cdot 0 + \ldots + \beta_{797591} \cdot 0$

↗ $= \beta_1 \cdot 0 + \beta_2 \cdot 0 + \ldots + \beta_{797591} \cdot 2$

Risk prediction models

- Generalized linear models:
  - Observations (x_i, y_i), i=1,…,n, with response y_i and covariate vector $x_i = (x_{i1}, \ldots, x_{ip})'$
  - Model for an exponential family response with known response function g

$$E(y_i|x_i) = g(\eta_i) = g(x_i'\beta)$$

- Cox proportional hazards model
  - Observations (t_i, δ_i, x_i), i=1,…,n, with observed time t_i, and δ_i, taking value 1 if an event occurred and 0 in case of censoring
  - Model for the hazard, i.e., the instantaneous risk,

$$h(t|x_i) = h_0(t)\exp(x_i'\beta)$$

  with unspecified baseline hazard $h_0(t)$
- Estimation of parameter vector β by via (partial) log-likelihood $l(\beta)$

Componentwise likelihood-based boosting

- Cox model $h(t|x_i) = h_0(t)\exp(x_i'\beta)$ with partial log-likelihood $l(\beta)$
- Start with estimate $\hat{\beta}_0 = (0, \ldots, 0)'$ and offset $\hat{\eta}_{i,0} = 0$
- For k=1,…,M, boosting steps (selected by cross-validation),
  1. Determine best covariate j^*:
     - Candidate models with penalized likelihood estimates $\hat{\alpha}_{j,k}$
     - (Penalized) score statistic
  2. Perform updates

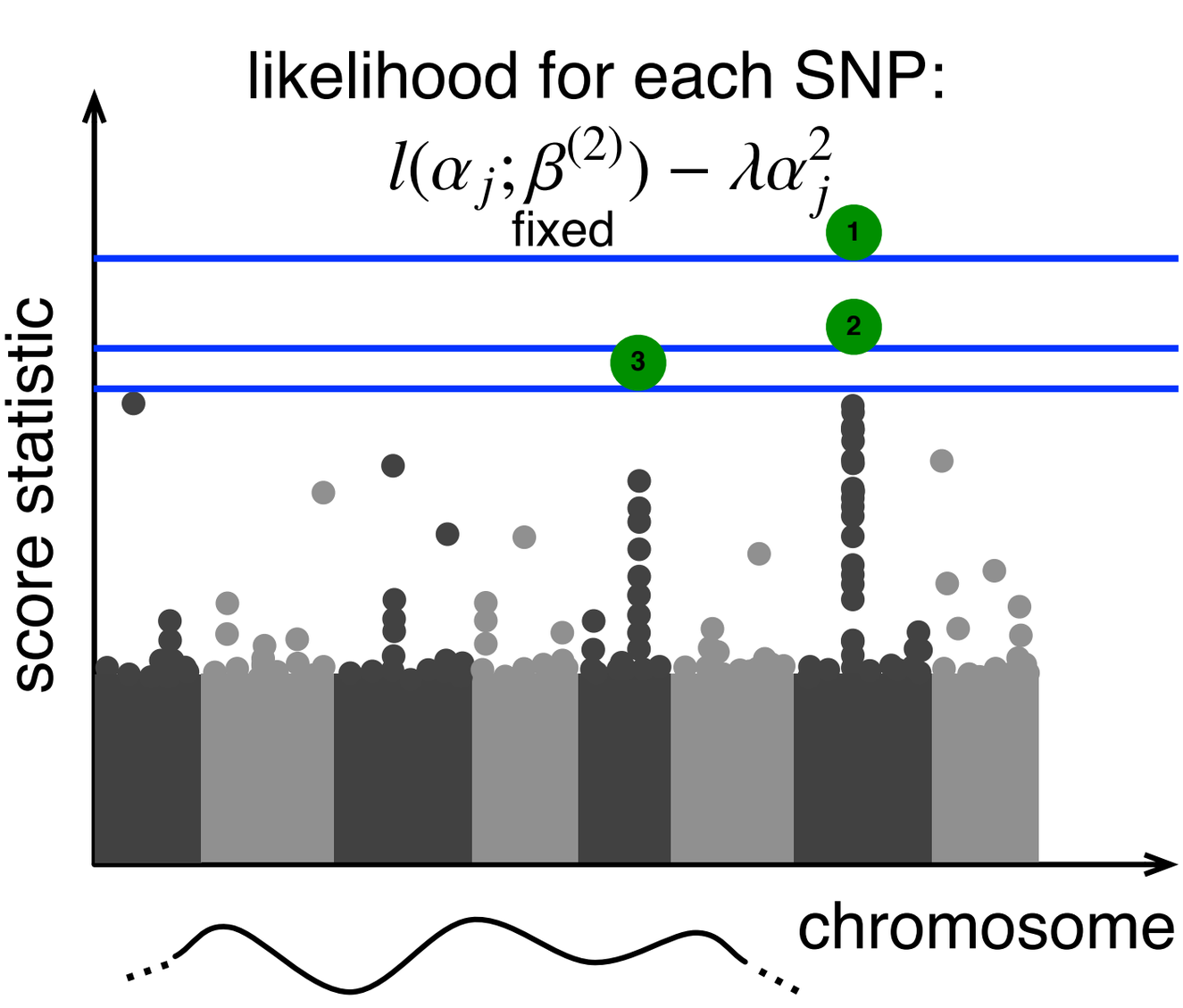
$$\hat{\beta}_{j,k} = \begin{cases} \hat{\beta}_{j,k-1} + \hat{\alpha}_{j,k} & \text{if } j = j^* \\ \hat{\beta}_{j,k-1} & \text{otherwise} \end{cases}$$

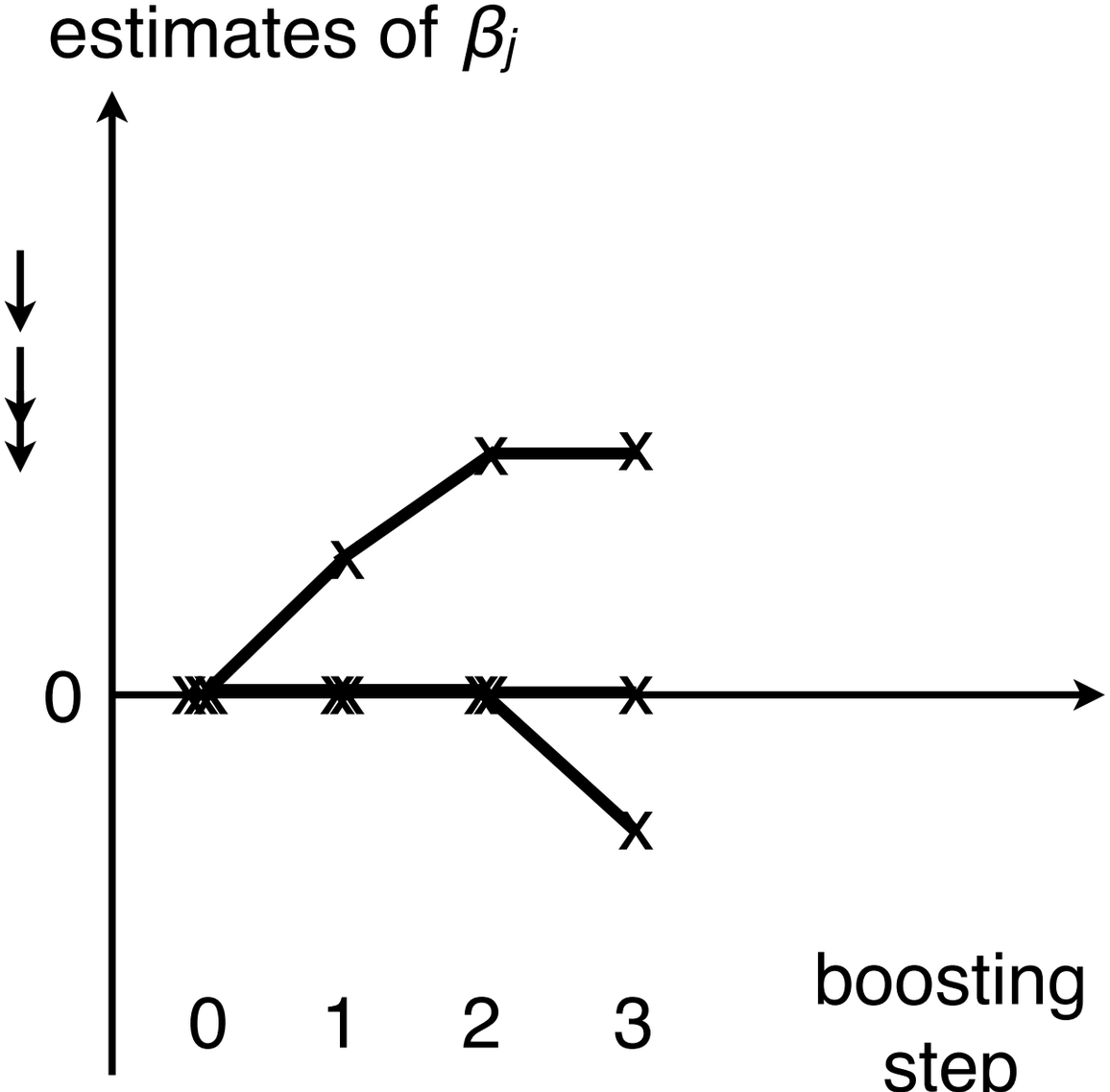
and

$$\hat{\eta}_{i,k} = x_i'\hat{\beta}_k$$

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Medizinische Biometrie,
Epidemiologie und Informatik

Componentwise likelihood-based boosting





JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Medizinische Biometrie,
Epidemiologie und Informatik

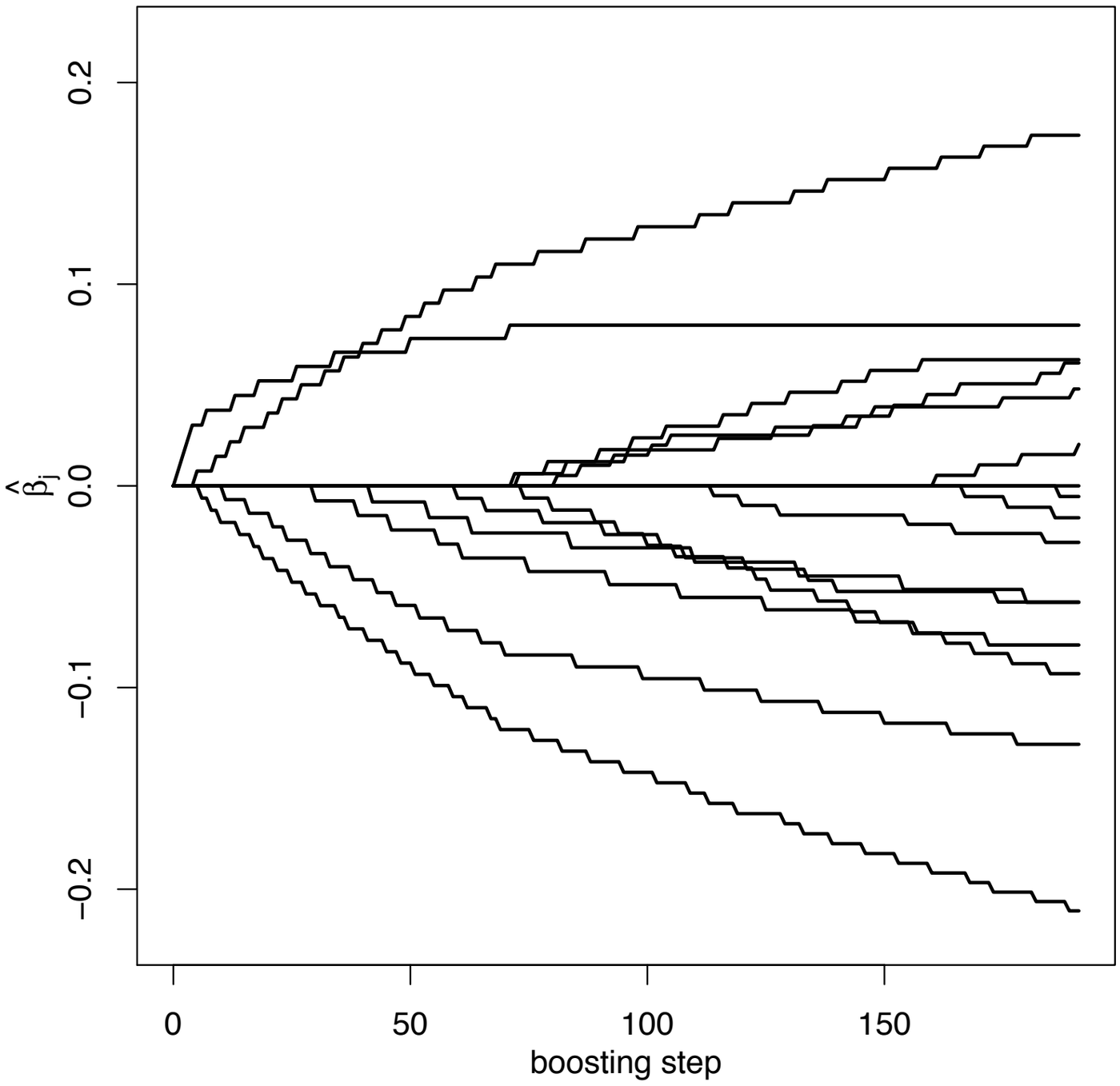
Adjusting for ...

established predictors

$$= \underbrace{\gamma_1 \cdot \text{age} + \gamma_2 \cdot \text{risk group} + \gamma_3 \cdot \text{treatment}}_{\textbf{un}\text{regularized}} + \underbrace{\beta_1 \cdot 2 + \beta_2 \cdot 2 + \ldots + \beta_{797591} \cdot 0}_{\text{regularized}}$$

Institut für Medizinische Biometrie,
Epidemiologie und Informatik

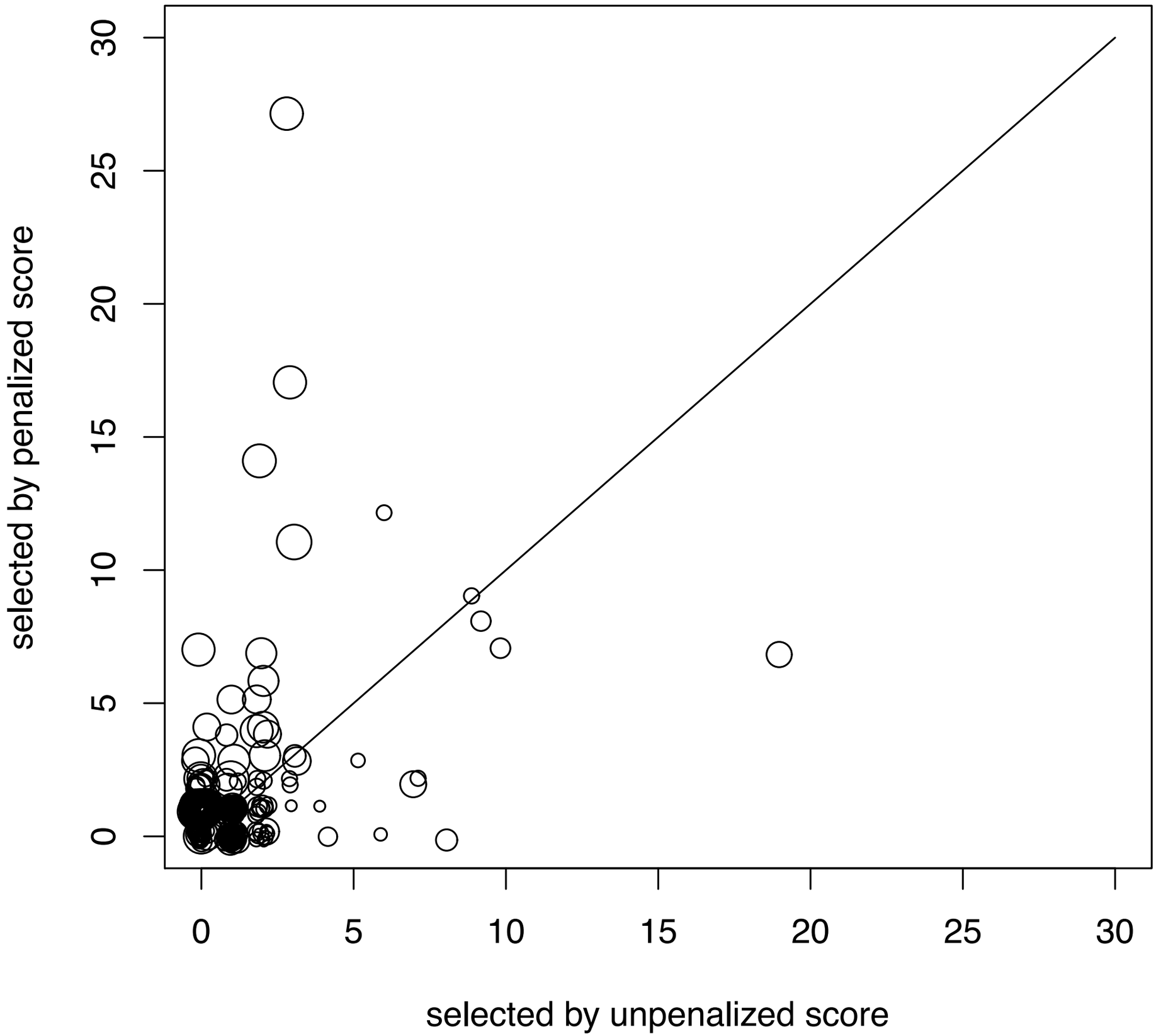
AML patients: coefficient paths



Choosing the score statistic ...

Penalized vs. Un-penalized

$$U_j^{(k)}(0)^2 / (I_j^{(k)}(0) + \lambda) \qquad U_j^{(k)}(0)^2 / I_j^{(k)}(0)$$


selected by penalized score
0
5
10
15
20
25
30
selected by unpenalized score

JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Medizinische Biometrie,
Epidemiologie und Informatik

Signature properties

	penalized	unpenalized
original signature	9	15
resampling Q50 / Q75	2 / 6	2 / 5
IF > 0 / > 10 / max	221 / 5 / 27	213 / 1 / 19

Bootstrap .632+ prediction error curves

- Apparent error

$$\overline{err}(t;\hat{r}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i(t) - \hat{r}(t|x_i))^2 W(t;\hat{G})$$

  overestimates performance

- Bootstrap cross-validation estimate

$$\widehat{Err}_{B0}(t,\hat{r}) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{b_0}\sum_{i\notin\mathcal{J}_b}(Y_i(t) - \hat{r}_b(t|x_i))^2 W(t,\hat{G})$$

  underestimates performance

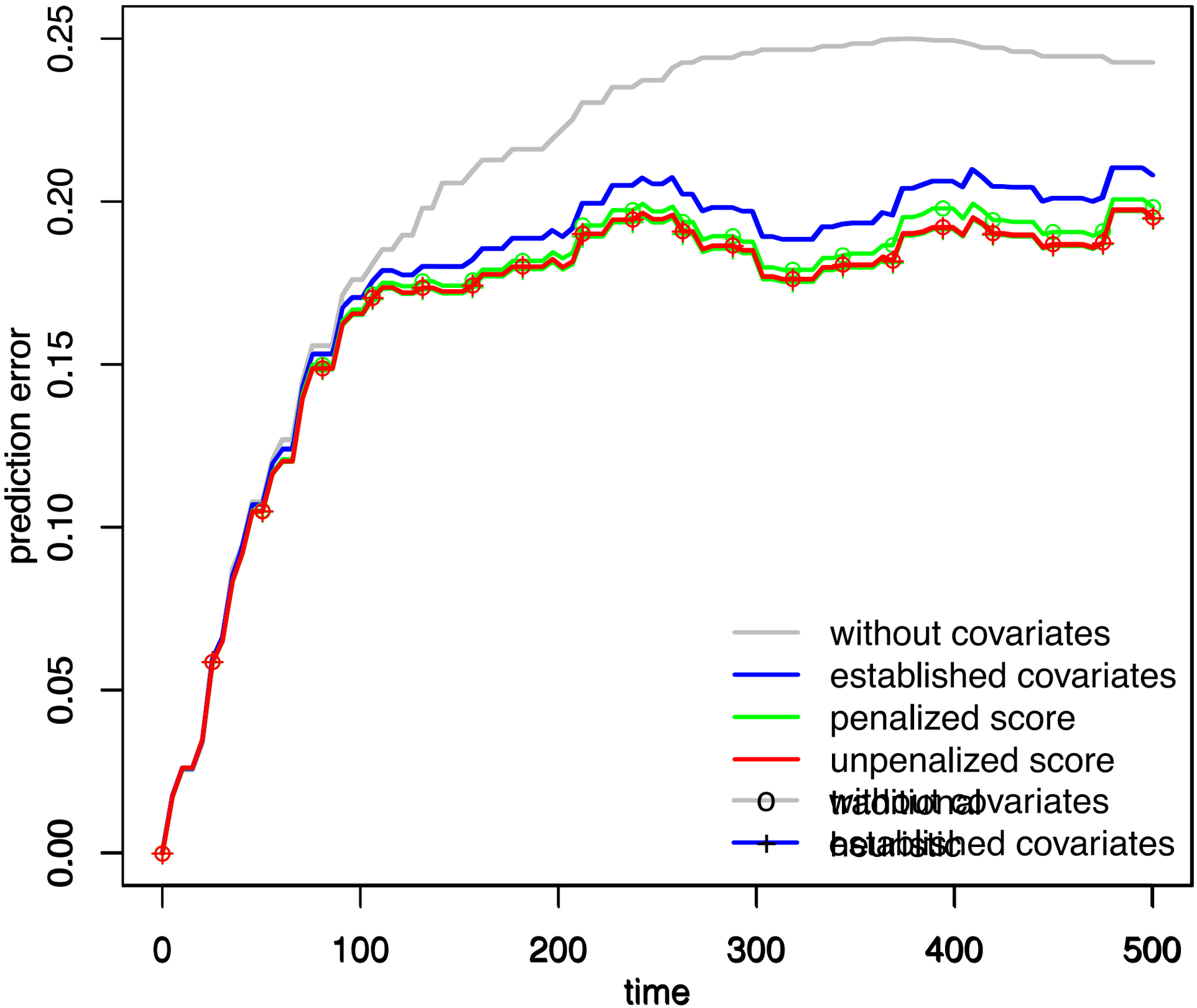
- Bootstrap .632+ estimate

$$\widehat{Err}_{.632+}(t,\hat{r}) = \left\{1-\omega(t)\right\}\overline{err}\,(t,\hat{r}) + \omega(t)\widehat{Err}_{B0}(t,\hat{r})$$

  adapts for overfitting potential

AML prediction error curves

Institut für Medizinische Biometrie,
Epidemiologie und Informatik



JG|U UNIVERSITÄTSmedizin. MAINZ
Institut für Medizinische Biometrie,
Epidemiologie und Informatik

Signature properties

SNP level:

	penalized	unpenalized
original signature	9	15
resampling Q50 / Q75	2 / 6	2 / 5
IF > 0 / > 10 / max	221 / 5 / 27	213 / 1 / 19

Gene level:

	penalized	unpenalized
original signature	17 (2/2)	19 (2/8)
resampling Q50 / Q75	4 / 9	3 / 7
IF > 0 / > 10 / max	249 / 11 / 43	252 / 4 / 29

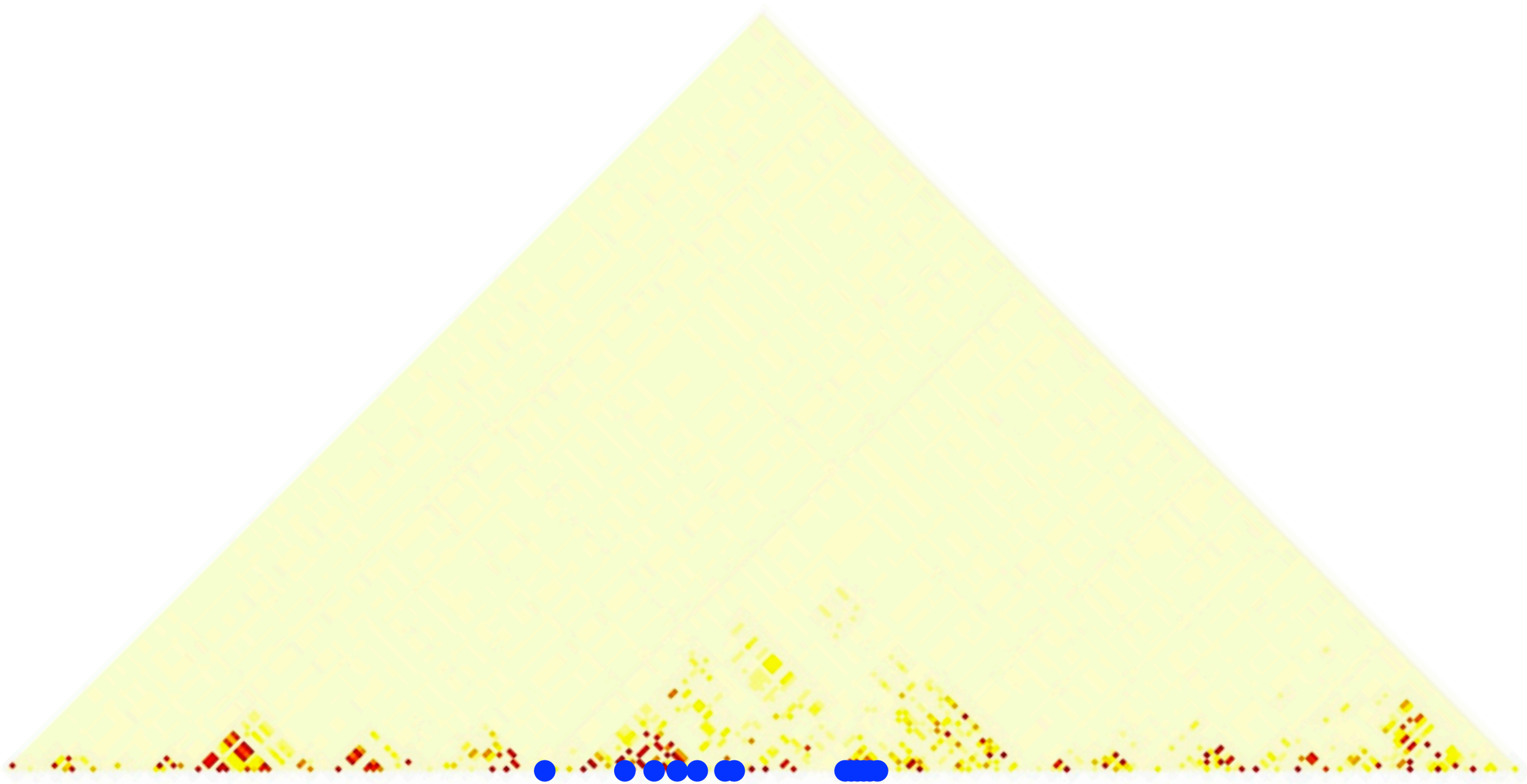
SNPs: univariate test-based strategy

- One Cox proportional hazards model per SNP:
  Test statistic from comparing models “with vs. without SNP”
- Inclusion frequencies: select top SNPs (same number as boosting)
- Gene level summary:
  - Maximum value of test statistic per gene
  - Null distribution/p-value via permutation

	componentwise boosting		univariate test-based		
Gene/SNP	IF gene	IF SNP	IF gene	IF SNP	p-value
FSTL4	48		37		0.005
SNP_A-1925137		1		1	
SNP_A-1851005		1		0	
SNP_A-2065481		0		1	
SNP_A-1894802		0		1	
SNP_A-4217770		0		1	
SNP_A-4265215		37		13	
SNP_A-1983529		7		10	
SNP_A-1909042 u		0		8	
SNP_A-4254590		13		18	
SNP_A-2096046 u		1		12	
SNP_A-2310208 u		0		9	
SNP_A-1849858 u		1		20	
SNP_A-2084916		0		1	
SNP_A-2038408		0		3	
CYP24A1	8		3		0.065
SNP_A-2111160 b		8		3	
SNP_A-2041818 b		7		3	

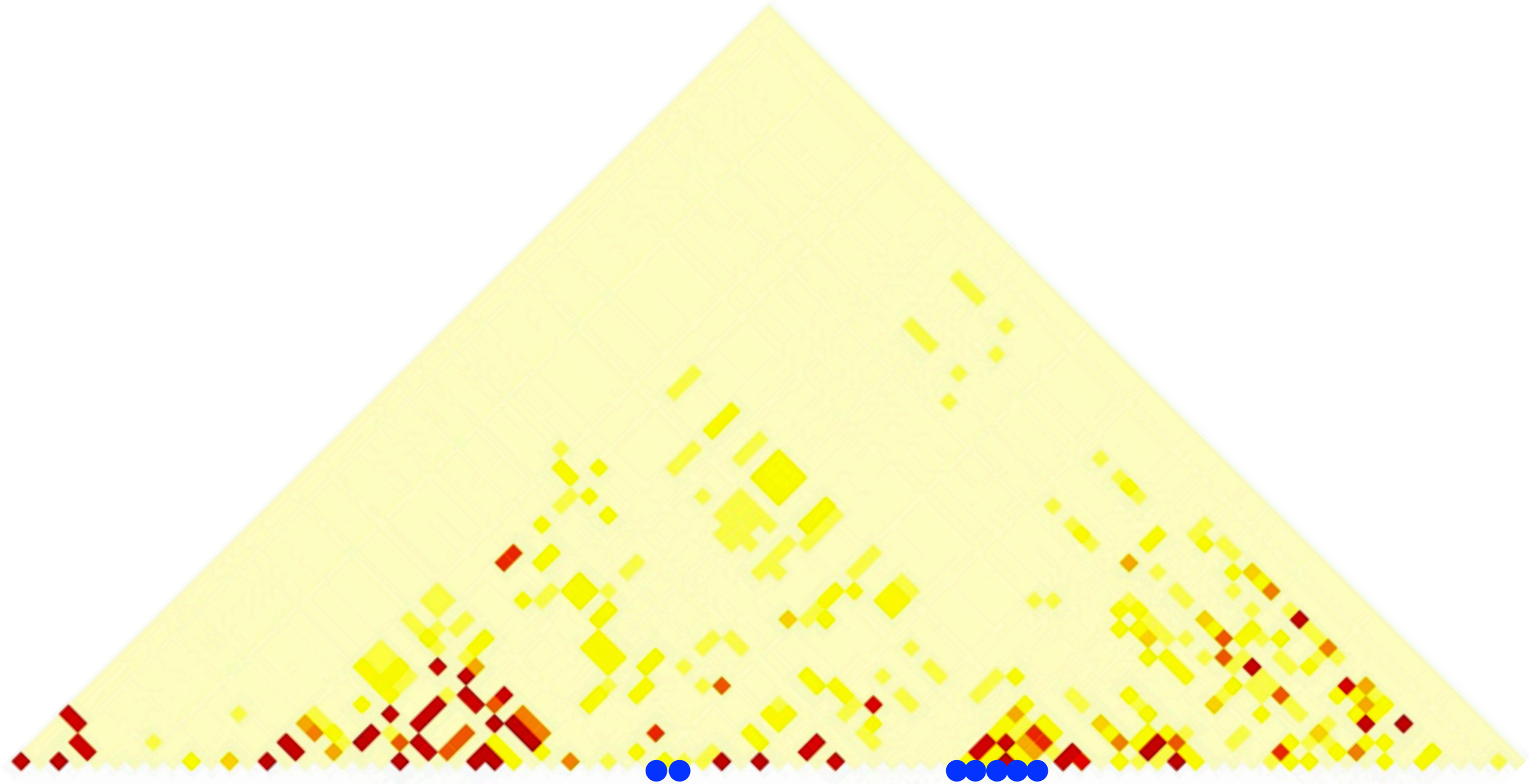
Institut für Medizinische Biometrie,
Epidemiologie und Informatik

Linkage disequilibrium for gene *FSTL4*



Institut für Medizinische Biometrie,
Epidemiologie und Informatik

Linkage disequilibrium for gene *FSTL4*



Summing up

- SNP data in clinical cohort setting:
  - risk prediction models
  - (only) identify SNP signature
- RNA-Seq:
  - different way of measuring gene expression
  - distribution problematic compared to microarray data
- Componentwise boosting:
  - monotone coefficient paths, compared to lasso
  - different ways for dealing with variance, standardization
  - pre-transformation useful for RNA-Seq data
- Compared to univariate strategy:
  - find at least some SNPs
  - increased stability

Institut für Medizinische Biometrie,
Epidemiologie und Informatik

Thanks to ...

Methods

Isabella Zwiener
IMBEI Mainz, Germany

Martin Schumacher
IMBI Freiburg, Germany

Stefanie Hieke
IMBI Freiburg

Axel Benner
DKFZ Heidelberg

Modeling for the AML data

Lars Bullinger
University Hospital of Ulm

Contact

Harald Binder

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

binderh@uni-mainz.de