Statistical techniques for linking high-dimensional molecular data to complex clinical endpoints

Harald Binder

[1]Freiburg Center for Data Analysis and Modeling, University of Freiburg, Germany
[2]Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg
supported by DFG Forschergruppe FOR 534

binderh@fdm.uni-freiburg.de

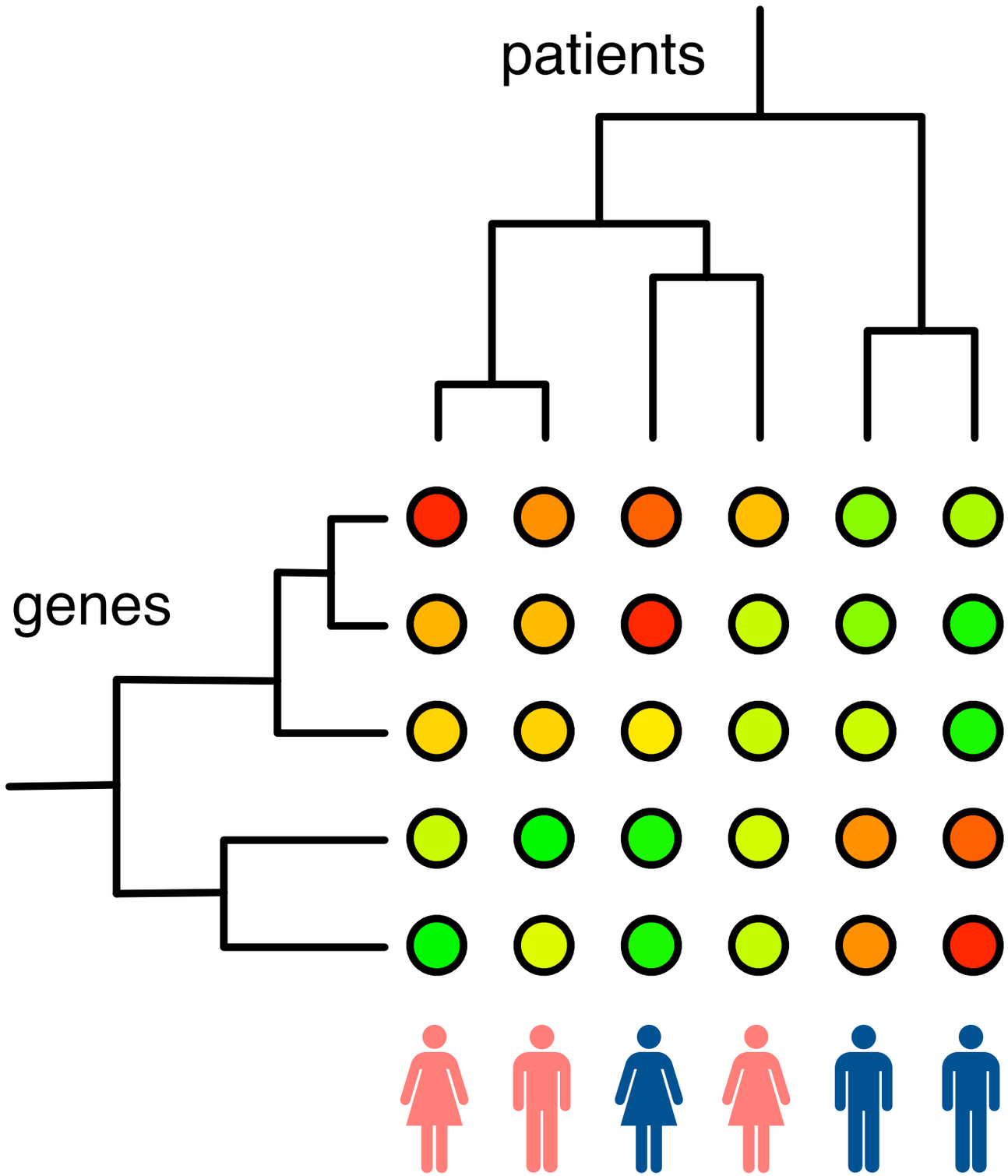
July 14, 2009

UNIVERSITÄTS KLINIKUM
FREIBURG

Project description

- Aim: Identify genes that are related to cardiovascular events
- Data:
  - 321 dialysis patients: 30 cardiovascular events, 71 deaths from other causes, 220 censored observations
  - 19 clinical covariates: e.g., age, sex, duration of dialysis, previous cardiovascular event
    - gene expression at baseline determined via 26323 microarray features
- Project head: Prof. Gerd Walz
- Lab partner: Thorsten Kurz
- Preprocessing: Clemens Kreutz

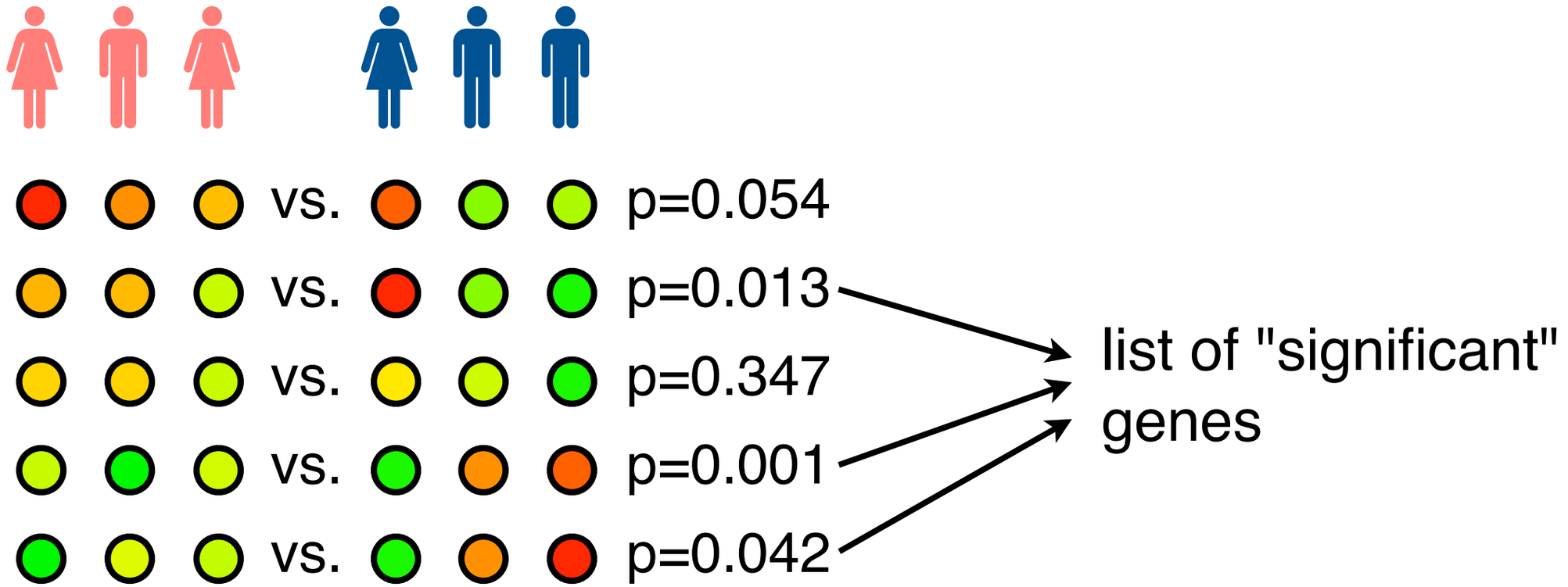
Strategy 1: Clustering



- Cluster genes w.r.t. similarity of expression across patients
- Cluster patients w.r.t. similarity of expression across genes
- Look for clusters where "affected" patients are overrepresented
- Main problem: Status of patients ("affected" vs. "unaffected") is not taken into account for clustering, i.e., **not optimizing for the right criterion**

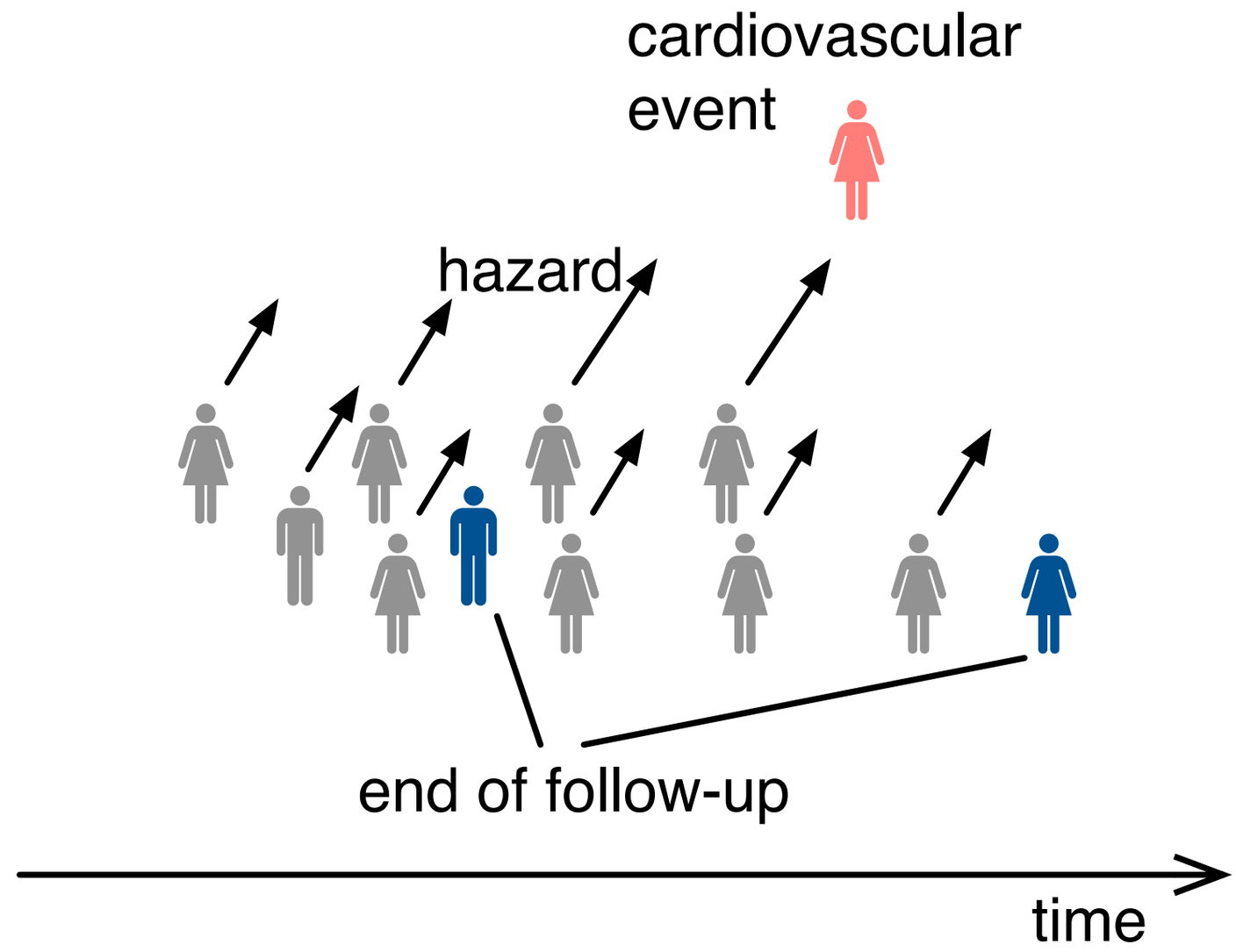
Strategy 2: Comparison of groups

- Contrast "affected" vs. "unaffected" for each gene



- Implicitly employs model "group → gene expression"
- Problems:
  - Model "gene expression → group membership" needed for judging potential for prediction of future cases
  - Does not fit if cohort instead of case-control design is employed

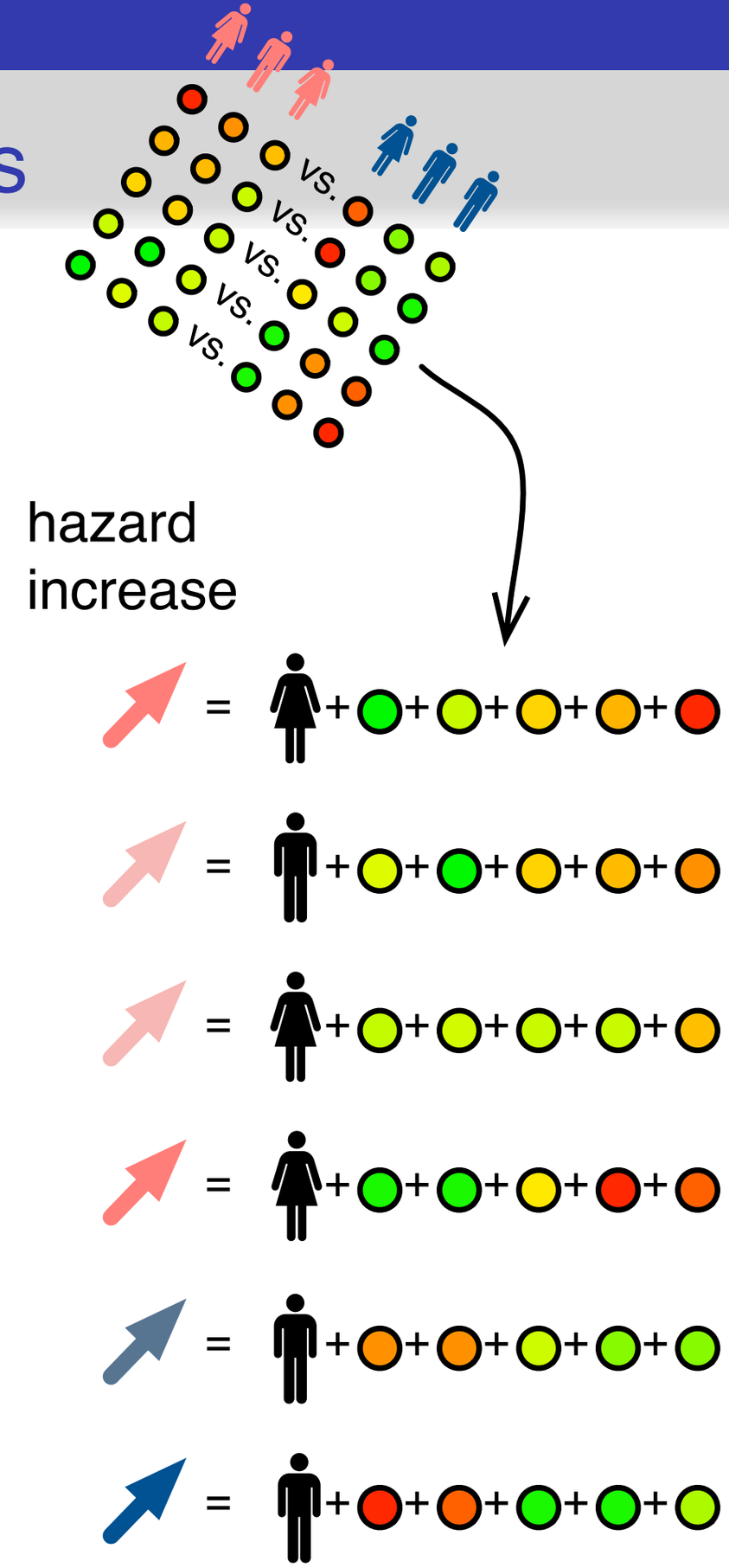
Structure in a simple time-to-event setting



- Patients experience events or are lost to follow-up at different times, i.e., they should not simply be grouped
- Take time into account by **modeling the hazard**, i.e., the instantaneous risk of having an event, as a function of time

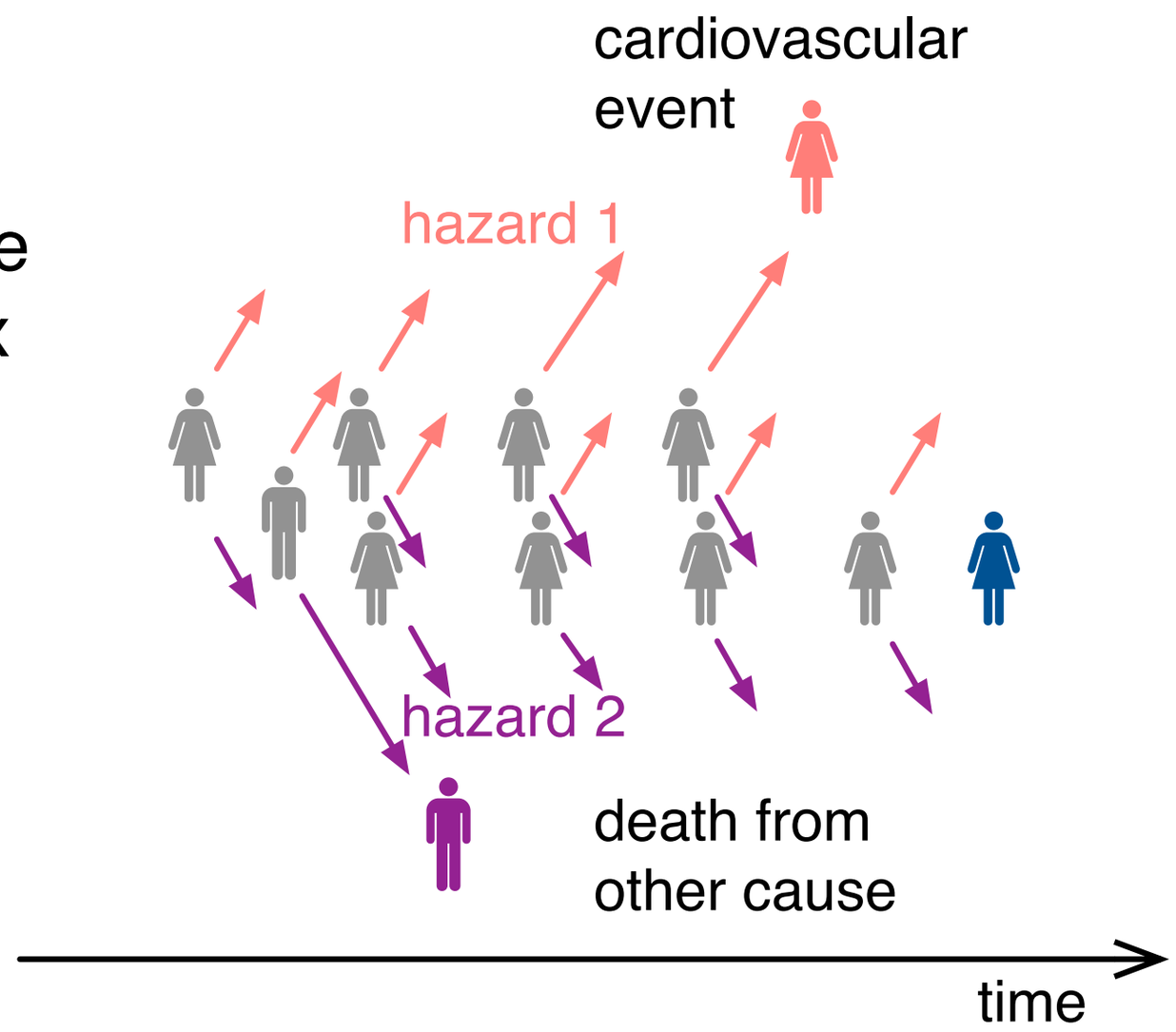
Strategy 3: Risk prediction models

- We fit risk prediction models that predict the **risk of having an event** up to a certain time for each individual
- **Cox proportional hazards model:** express the hazard via a linear combination of clinical covariates (e.g., sex) and gene expression measurements. i.e., "gene expression → hazard"
- The contribution, i.e., the **importance**, of a single clinical covariate or gene expression measurement is **expressed via** a **regression coefficient**



Complex endpoints: competing risks

- Competing events that might also be connected to some gene expression patterns cannot be ignored → consider all hazards
- **Comprehensive model:** Fine&Gray model for the proportion of cardio-vascular events (cumulative incidence) extends the Cox model for adequately considering competing events

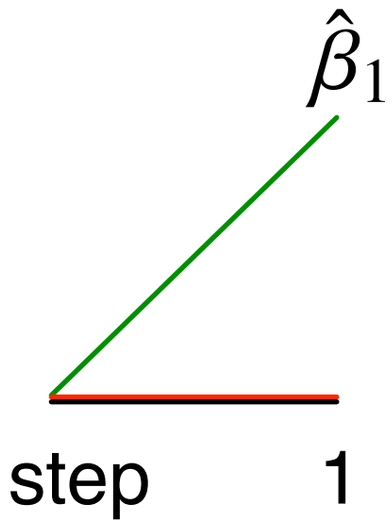


Sparse risk prediction models

- Contribution of each gene expression value quantified via a regression coefficient, i.e., **regression coefficient equal to zero** means that the corresponding gene is **not part of the model**
- Sparse risk prediction models, i.e., models with a small number of non-zero regression coefficients, provide a short list of important genes → **We get also a gene list** (in addition to predictions)
- Employing a large number of small steps for building up regression coefficients:

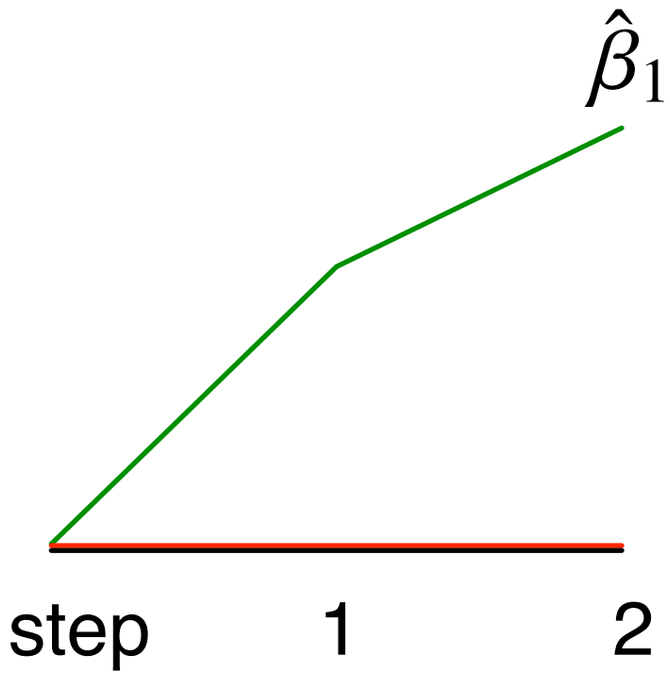
Sparse risk prediction models

- Contribution of each gene expression value quantified via a regression coefficient, i.e., **regression coefficient equal to zero** means that the corresponding gene is **not part of the model**
- Sparse risk prediction models, i.e., models with a small number of non-zero regression coefficients, provide a short list of important genes → **We get also a gene list** (in addition to predictions)
- Employing a large number of small steps for building up regression coefficients:



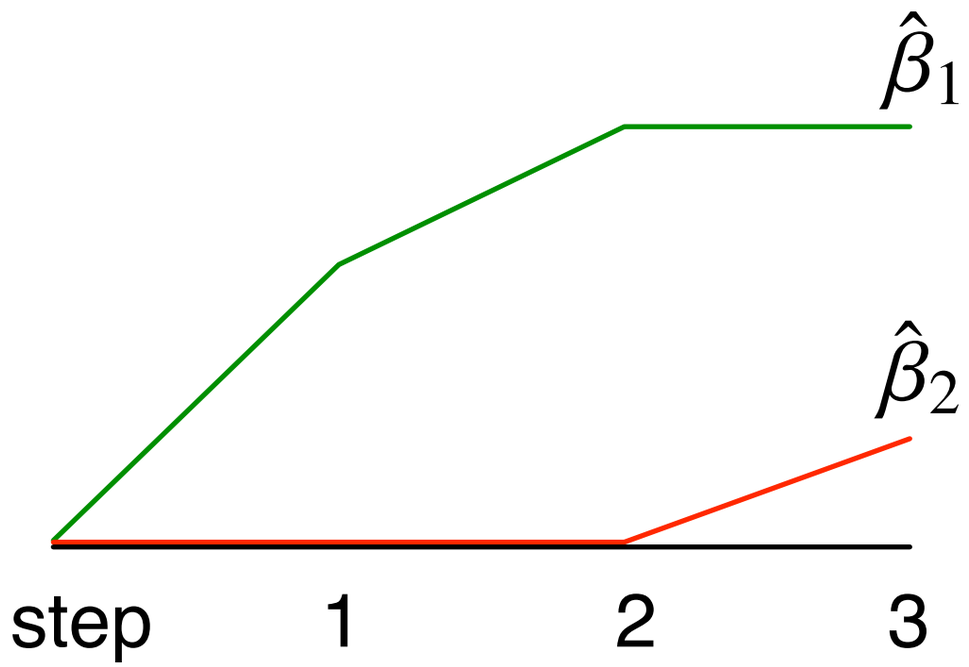
Sparse risk prediction models

- Contribution of each gene expression value quantified via a regression coefficient, i.e., **regression coefficient equal to zero** means that the corresponding gene is **not part of the model**
- Sparse risk prediction models, i.e., models with a small number of non-zero regression coefficients, provide a short list of important genes → **We get also a gene list** (in addition to predictions)
- Employing a large number of small steps for building up regression coefficients:



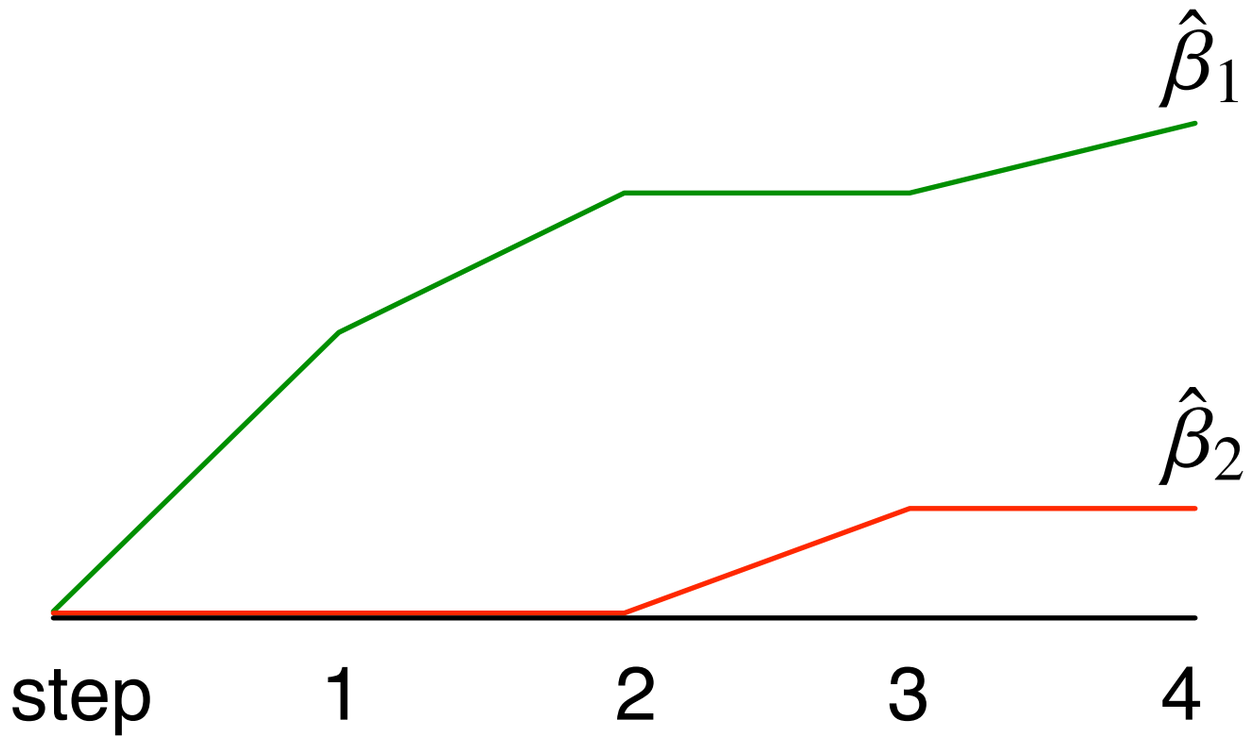
Sparse risk prediction models

- Contribution of each gene expression value quantified via a regression coefficient, i.e., **regression coefficient equal to zero** means that the corresponding gene is **not part of the model**
- Sparse risk prediction models, i.e., models with a small number of non-zero regression coefficients, provide a short list of important genes → **We get also a gene list** (in addition to predictions)
- Employing a large number of small steps for building up regression coefficients:

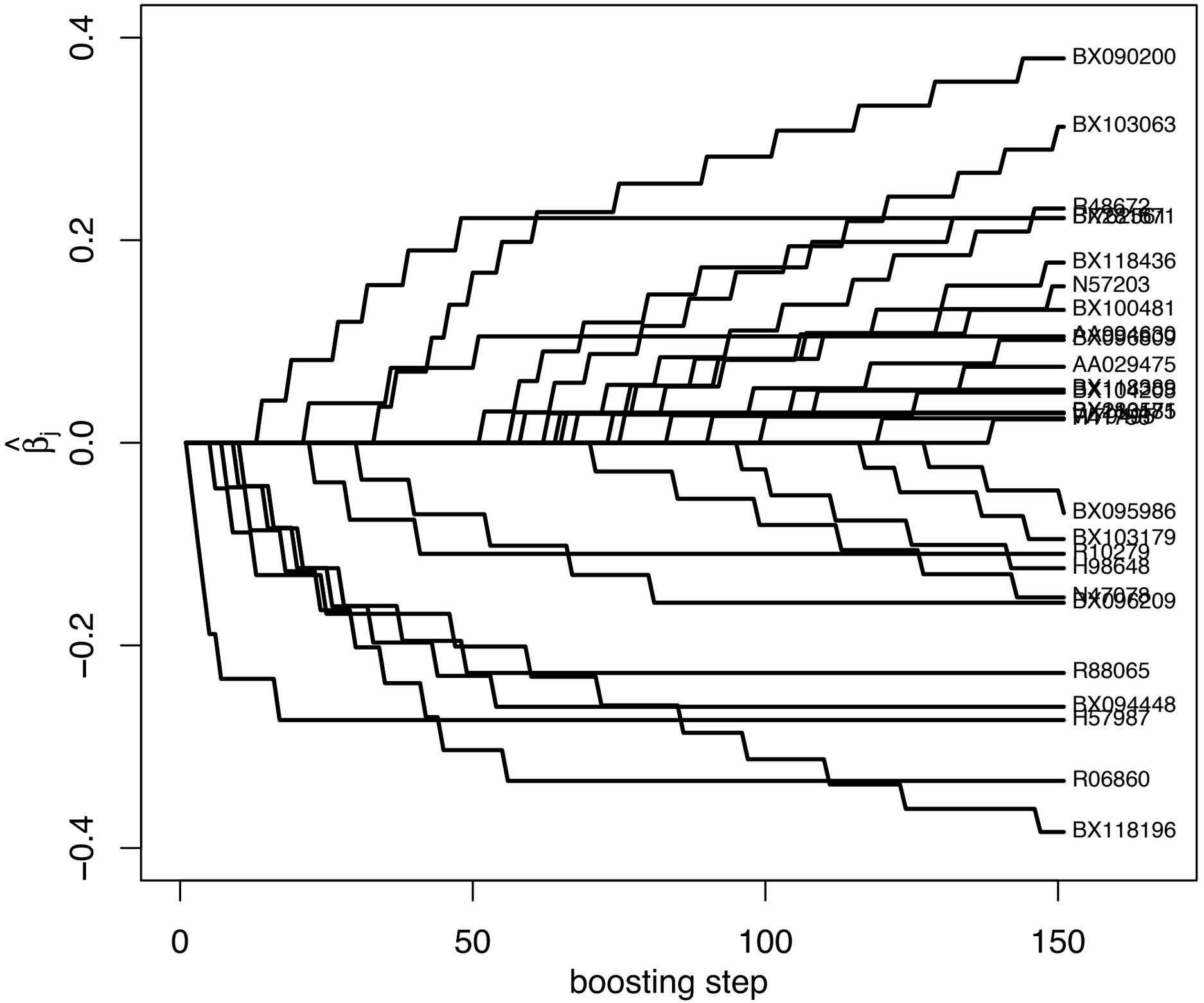


Sparse risk prediction models

- Contribution of each gene expression value quantified via a regression coefficient, i.e., **regression coefficient equal to zero** means that the corresponding gene is **not part of the model**
- Sparse risk prediction models, i.e., models with a small number of non-zero regression coefficients, provide a short list of important genes → **We get also a gene list** (in addition to predictions)
- Employing a large number of small steps for building up regression coefficients:

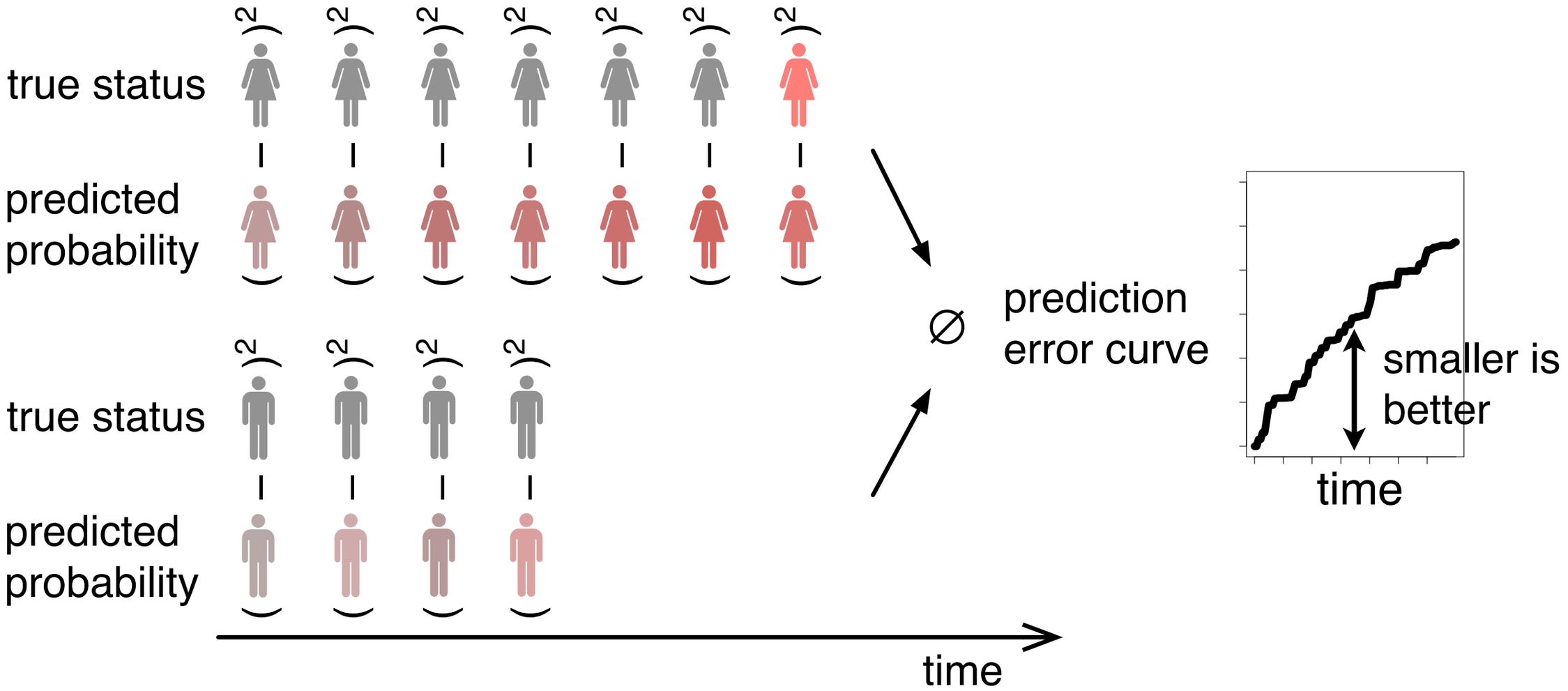


Cardiovascular events: regression coefficients



How to evaluate risk prediction models?

- For gene lists from group comparisons: p-values for evaluating performance
- Risk prediction models are built for predicting future observations
  → **prediction error for performance evaluation**



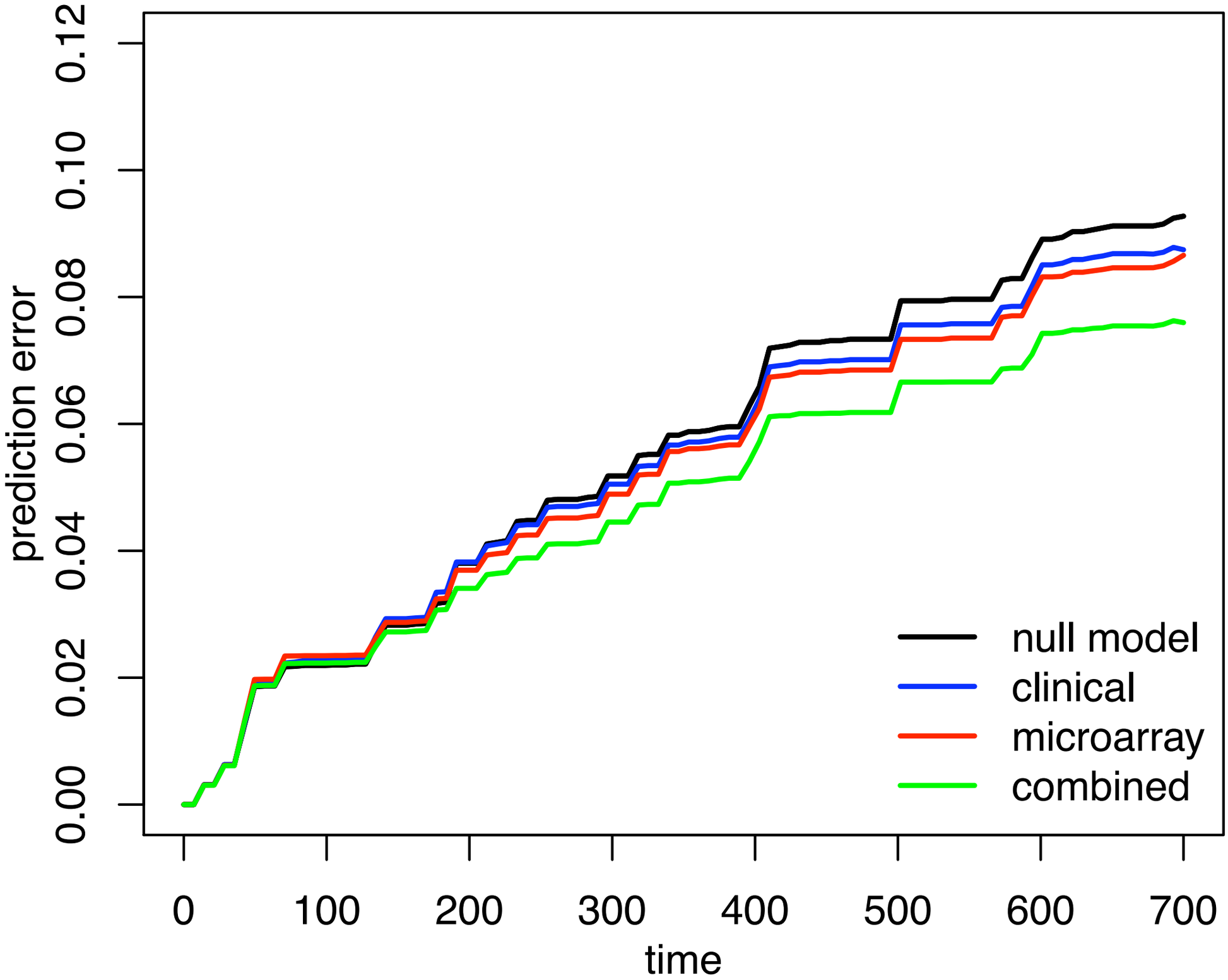
Pitfalls in model evaluation

- With thousands of gene expression measurements "perfect" prediction can always be obtained on the data that was used for fitting a risk prediction model
  → not useful for judging prediction performance in new data

  Alternatives:

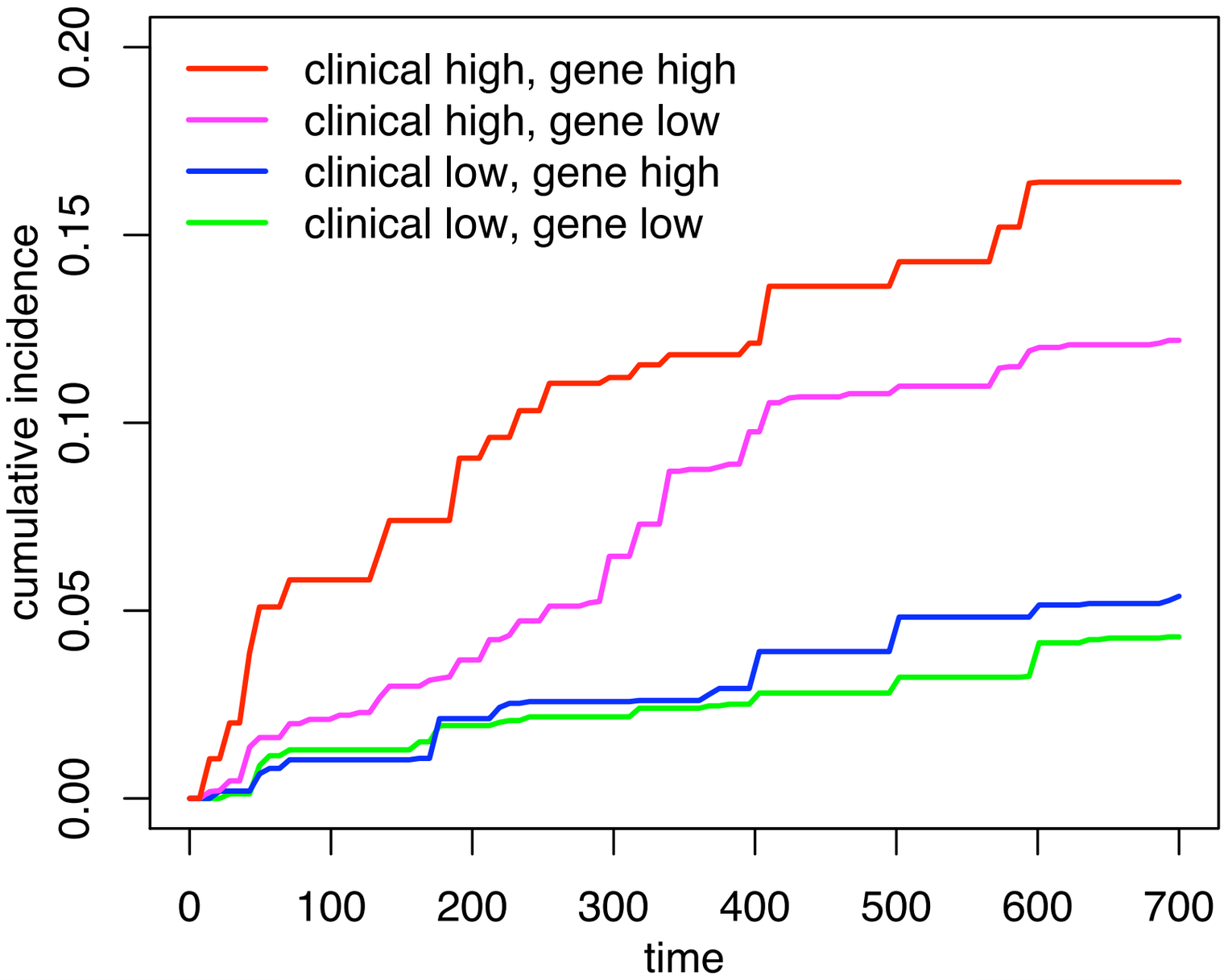
  - Set aside test data → too expensive
  - **Bootstrap:** Repeatedly, generate "new" data by randomly drawing observations, and evaluate on left-out observations
    Important: all model building steps in each bootstrap sample
- Even with an unbiased prediction error estimate it is important to employ **adequate performance references**:
  - Null model that does not employ any covariate information
  - Purely clinical model

Predicting cardiovascular events



We even can get groups...

- A risk prediction model assigns each patient a predicted probability of experiencing an event
- If needed, this can be categorized for obtaining risk groups:



Incorporating pathway information

- There are several databases that provide information on relations of genes. E.g., the KEGG pathway database describes relations of genes in pathways
- Gene lists from sparse risk prediction models often incorporate only few genes from a pathway, even when the whole pathway has an effect
- We developed techniques for incorporating external pathway information, for recovering larger parts of pathways, guided by prediction performance
- We also can infer connection signs of gene relations at the same time while estimating risk prediction models

Summary / Future research

- Thinking in terms of risk prediction models allows for linking high-dimensional molecular data to complex clinical endpoints
- Models are evaluated via prediction performance, i.e., added value of molecular data can be easily quantified
- With sparse techniques, short lists of informative genes are obtained
- Additional knowledge, e.g., pathway information, can be incorporated
- **Future research:**
  - Combining data from several sources, e.g., SNPs and gene expression, mRNA and microRNA, gene expression and protein mass spectra
  - More complex clinical endpoints, e.g., multistate models
  - Time-dependent measurements

References (from our group)

- Binder, H., Allignol, A., Schumacher, M., and Beyersmann, J. (2009). Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*, 25(7):890-896.
- Binder, H. and Schumacher, M. (2009). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics*, 10:18.
- Porzelius, C., Binder, H., and Schumacher, M. (2009). Parallelized prediction error estimation for evaluation of high-dimensional models. *Bioinformatics*, 25(6):827-829.
- Binder, H. and Schumacher, M. (2008a). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 9:14.
- Binder, H. and Schumacher, M. (2008b). Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. *Statistical Applications in Genetics and Molecular Biology*, 7(1):Article 12.
- Schumacher, M., Binder, H., and Gerds, T. A. (2007). Assessment of survival prediction models based on microarray data. *Bioinformatics*, 23(14):1768-1774.