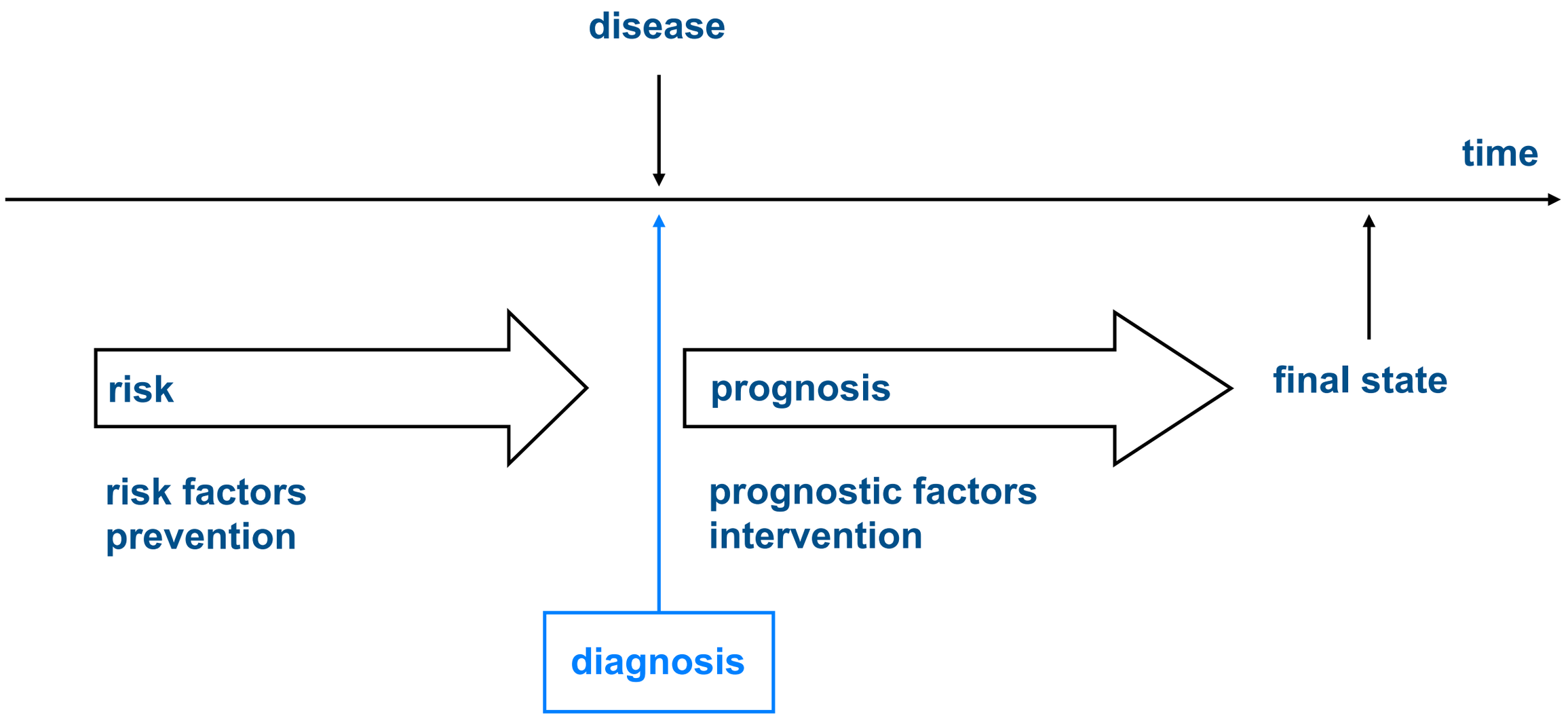
JG|U UNIVERSITÄTSmedizin. MAINZ

# Generalized linear models and high-dimensional applications

Harald Binder




**epidemiological studies, e.g. cross-sectional and cohort studies**

**screening, diagnostic studies**

**intervention studies, studies on prognostic factors**

UNIVERSITÄTSmedizin. MAINZ
Institut für Medizinische Biometrie, Epidemiologie und Informatik

# Dangers in observational studies

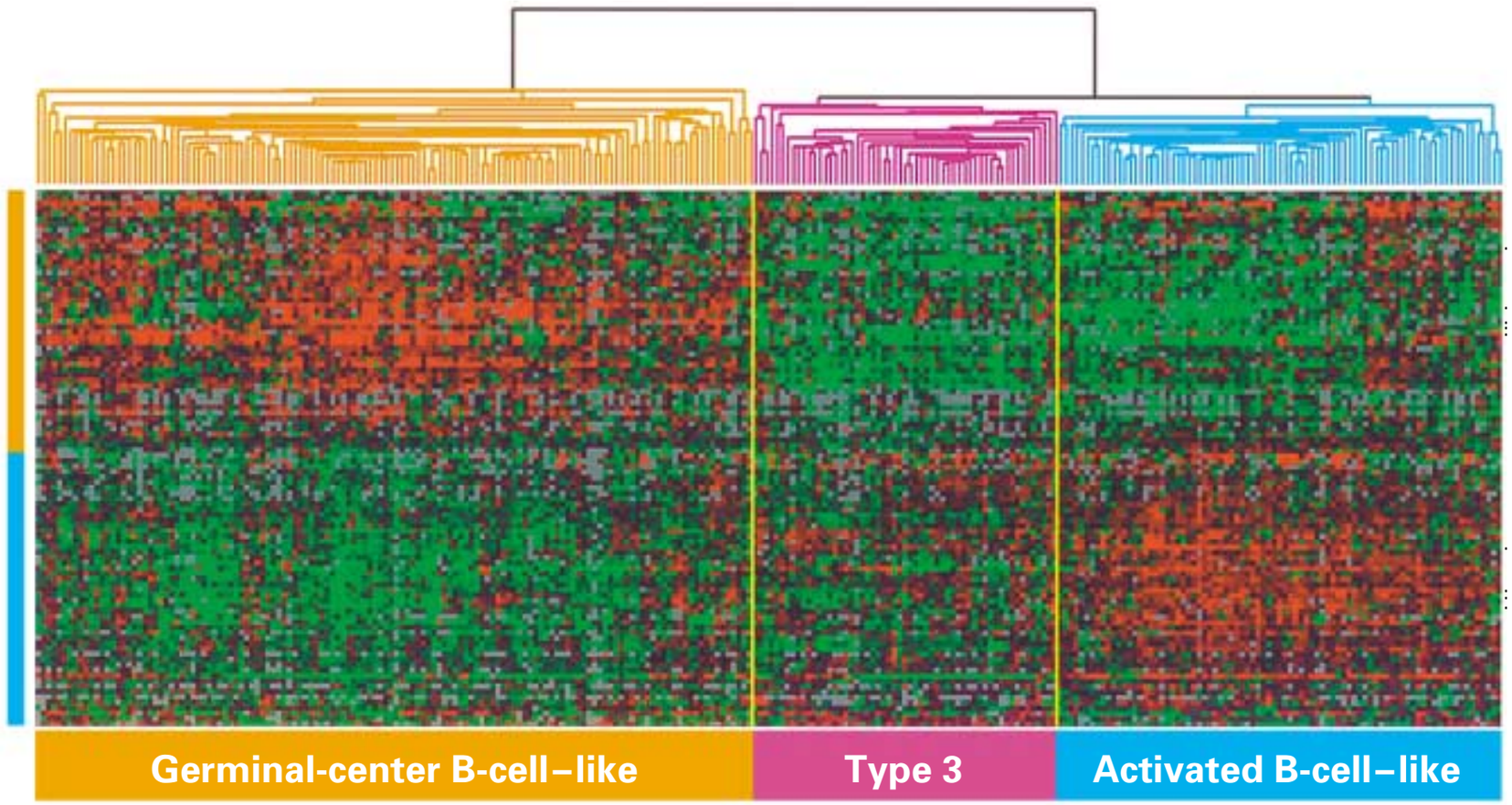
## Simpson's paradox

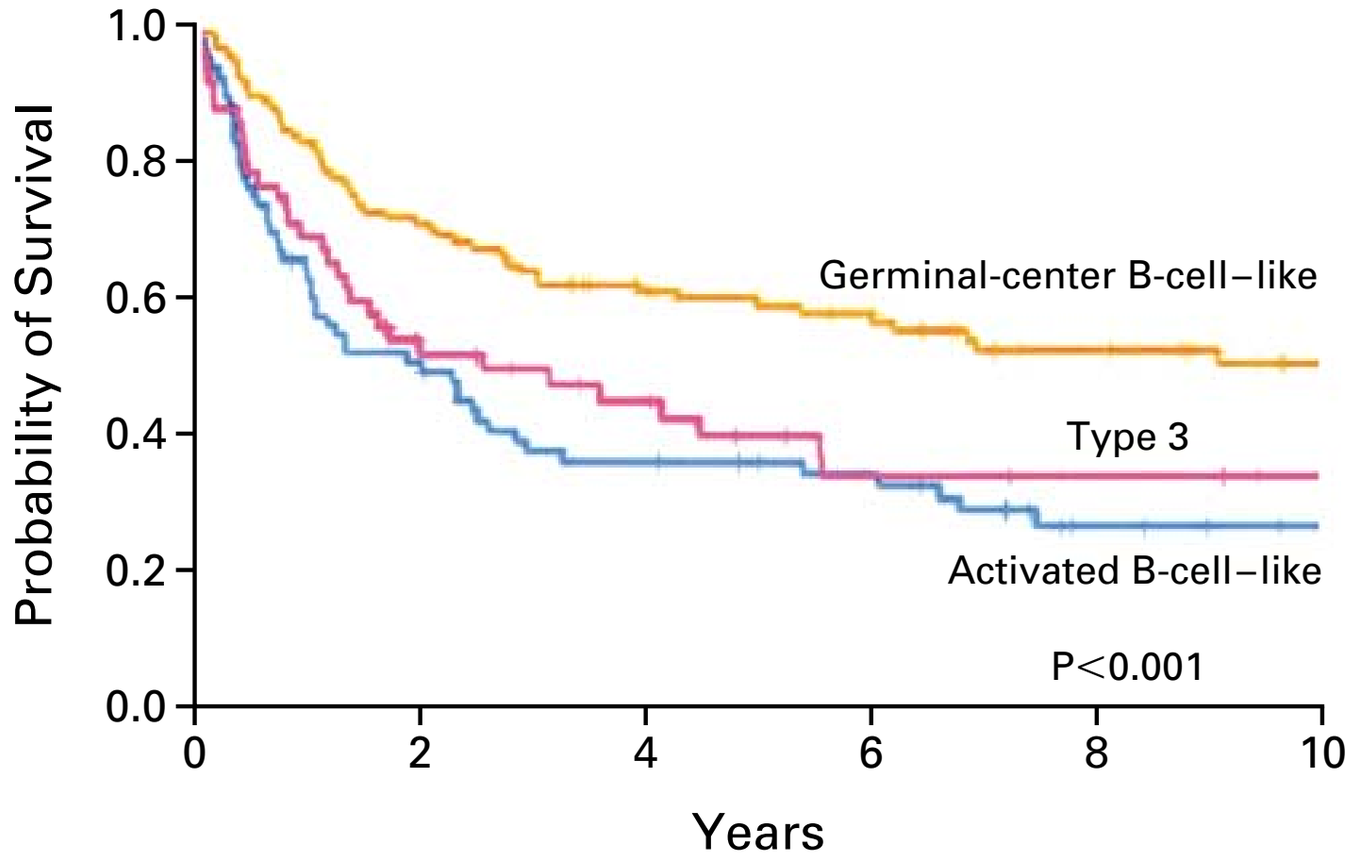
| | therapy | alive | dead | all | |
|---|---|---|---|---|---|
| male | **A** | 30 | 270 | 300 | odds ratio > 1 |
| | **B** | 6 | 144 | 150 | |
| | **all** | 36 | 414 | 450 | |
| female | **A** | 30 | 30 | 60 | odds ratio > 1 |
| | **B** | 130 | 260 | 390 | |
| | **all** | 160 | 290 | 450 | |
| all | **A** | 60 | 300 | 360 | odds ratio < 1 |
| | **B** | 136 | 404 | 540 | |
| | **all** | 196 | 704 | 900 | |

# Application example: Rosenwald DLBCL study

- Prognostic study (Rosenwald et al., 2002) with retrospective collection of tumor-biopsy specimens and clinical data
- 240 patients with untreated diffuse large-B-cell lymphoma (DLBCL)
- 138 deaths, 5-year overall survival: 48%
- Existing clinical predictor: International Prognostic Index (IPI), which is a summary of 5 clinical features
- Lymphochip cDNA microarray technology resulting in p=7399 microarray covariates
- Questions:
  - Do the genes contain "interesting" information?
  - What is the added value of the microarray information
  - Which ones are "important" genes?

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

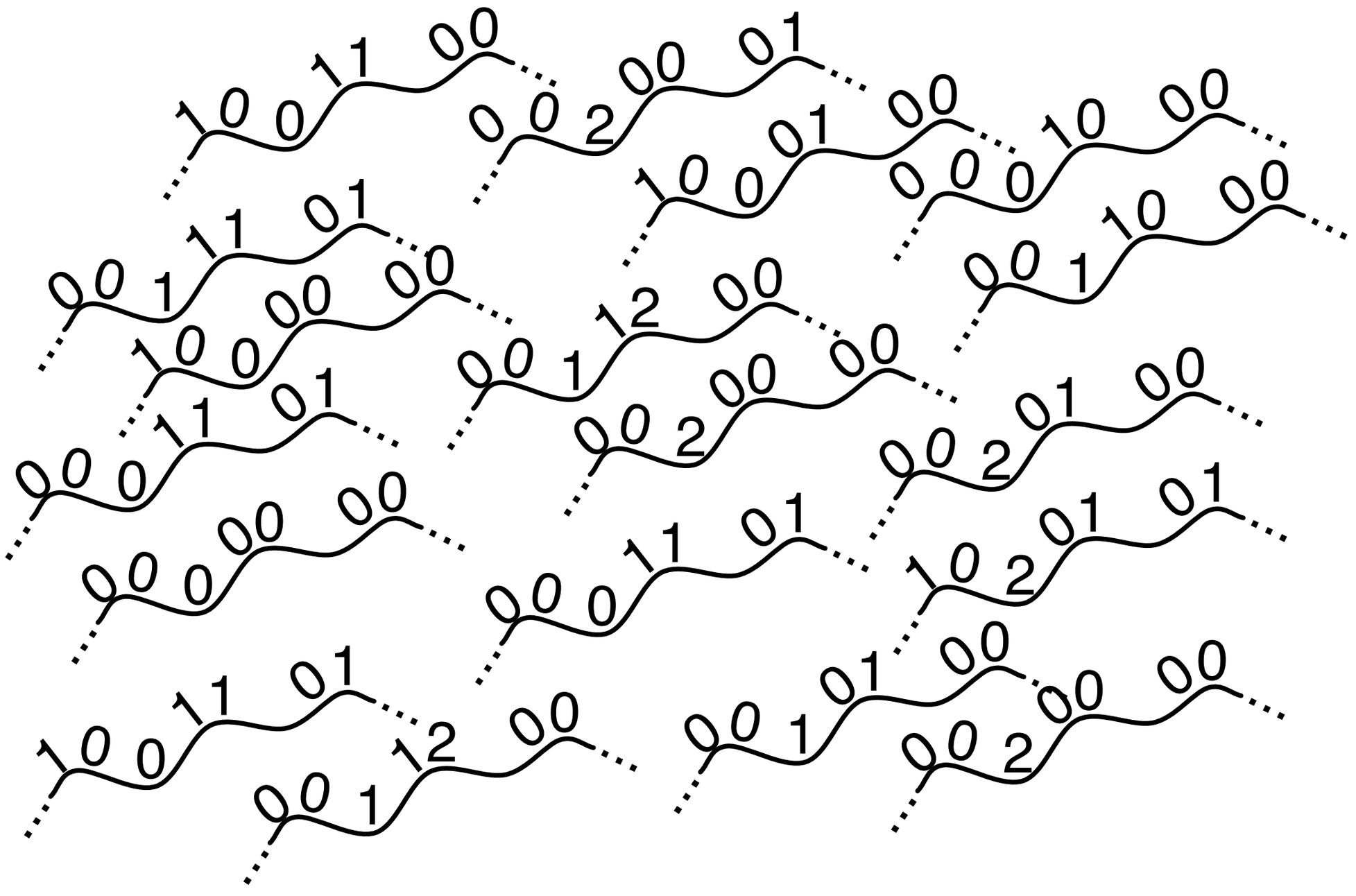



Rosenwald et. al (2002)

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

# Single nucleotide polymorphisms (SNPs)

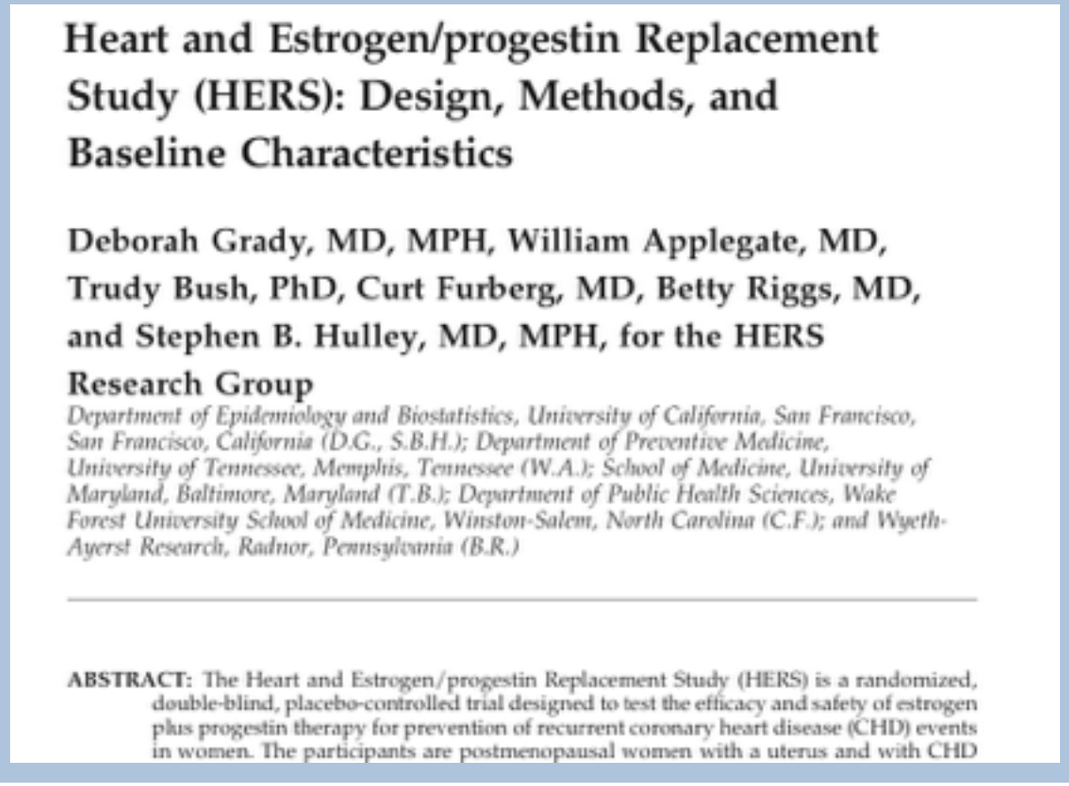


# Regressions models: basics

- Regression models ...
  - relate one ore more potential influencing variables (also: covariates, independent variables) to a target (also: response, endpoint, dependent variable)
  - allow for different types of endpoints (continuous, binary, time-to-event), requiring specific types of models
  - require specification of the structure of the relations between covariates and endpoints via a regression equation
- Simple case: linear regression for a continuous endpoint
  - continuous endpoint
  - additive, linear effect of the covariates
  - most regression modeling techniques are an extension of this

# Example for a continuous endpoint

- Heart and estrogen/progestin replacement Study (HERS)
- Sample: observational data of 2028 women
- Endpoint: glucose level
- Questions: effect of sport ("yes = 3 an more time per week“ / no = less than 3 times per week")



Heart and Estrogen/progestin Replacement Study (HERS): Design, Methods, and Baseline Characteristics

Deborah Grady, MD, MPH, William Applegate, MD, Trudy Bush, PhD, Curt Furberg, MD, Betty Riggs, MD, and Stephen B. Hulley, MD, MPH, for the HERS Research Group

*Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California (D.G., S.B.H.); Department of Preventive Medicine, University of Tennessee, Memphis, Tennessee (W.A.); School of Medicine, University of Maryland, Baltimore, Maryland (T.B.); Department of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina (C.F.); and Wyeth-Ayerst Research, Radnor, Pennsylvania (B.R.)*

**ABSTRACT:** The Heart and Estrogen/progestin Replacement Study (HERS) is a randomized, double-blind, placebo-controlled trial designed to test the efficacy and safety of estrogen plus progestin therapy for prevention of recurrent coronary heart disease (CHD) events in women. The participants are postmenopausal women with a uterus and with CHD

# t-test

- Theoretical quantity for the population of all women: glucose level for women without sport $\mu_1$ and with sport $\mu_2$
- Estimation of both parameters using the mean values in the sample: $\bar{x}_1$ (=97.4) and $\bar{x}_2$ (=95.6)
- For judging whether the difference indicates a difference in the population: standardized difference of the means as a test statistic
- $t_{(n\text{-}1)}$-distribution for:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\hat{SE}(\bar{x}_1 - \bar{x}_2)}$$

# Written as a regression model

- glucose level = basis glucose level
  + effect of other factors (incl. sport)
  + effect of random factors
- Variable "sport" taking values 0 (for "no") and 1 (for "yes")
- Model *M1*:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

- $\beta_0$ (intercept): glucose level for "no sport"
- $\beta_1$: difference in glucose level for "sport"
- $\varepsilon$: normally distributed error term

# Estimating the regression parameters

- Regression parameters $\beta_j$ are unknown; estimates $\hat{\beta}_j$ based on data by *fitting* a regression model
- Standard error $SE(\hat{\beta}_j)$ allows to determine confidence intervals
- Statistical test for $H_0$: $\beta_j = 0$ vs. $H_1$: $\beta_j \neq 0$ using statistic

$$T = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

# Examples: HERS

- Estimates for model *M1*:

| | estimate | SE | T | p-value |
|---|---|---|---|---|
| **$\beta_0$ (intercept)** | 97.38 | 0.28 | 346.00 | < 0.00001 |
| **$\beta_1$ (sport)** | -1.74 | 0.44 | -3.99 | 0.00007 |

- For comparison: t-test
  T = (95.6 - 97.4) / 0.437 = -4.023

# Taking additional factors into account

- Women with more sport potentially also younger, different level of alcohol consumption, or potentially also a different body mass index (BMI)
- Adjusting for potential confound by entering them into the regression model,

$$\text{glucose level} = \beta_0 + \beta_{1\,(\text{sport})} \cdot \text{"sport 0/1"} + \beta_{2\,(\text{age})} \cdot \text{age} + \beta_{3\,(\text{alcohol})} \cdot \text{"alcohol consumption 0/1"} + \beta_{4\,(\text{BMI})} \cdot \text{BMI}$$

- Interpreting the effect of "sport 0/1" while holding the other factors constant

# Example: HERS

- Estimates for model *M2*:

| | estimate | SE | T | p-value |
|---|---|---|---|---|
| $\beta_0$ (intercept) | 78.96 | 2.59 | 30.45 | < 0.00001 |
| $\beta_1$ (sport) | -0.95 | 0.43 | -2.22 | 0.0267 |
| $\beta_2$ (age) | 0.06 | 0.03 | 2.02 | 0.0431 |
| $\beta_3$ (alcohol) | 0.68 | 0.52 | 1.61 | 0.1071 |
| $\beta_4$ (BMI) | 0.49 | 0.04 | 11.77 | < 0.00001 |

- For comparison: model *M1*:

| | estimate | SE | T | p-value |
|---|---|---|---|---|
| $\beta_0$ (intercept) | 97.38 | 0.28 | 346.00 | < 0.00001 |
| $\beta_1$ (sport) | -1.74 | 0.44 | -3.99 | 0.00007 |

# Prediction

- Both models can be used for obtaining a prediction (or index) for the glucose level:

  from *M1*: $= 97.38 - 1.74 \cdot \text{sport}$

  from *M2*: $= 78.96 - 0.95 \cdot \text{sport} + 0.06 \cdot \text{age} + 0.68 \cdot \text{alcohol} + 0.49 \cdot \text{BMI}$

# Odds Ratio

| | D+ | D− | all |
|---|---|---|---|
| E+ | 30 | 270 | 300 |
| E− | 6 | 144 | 150 |
| all | 36 | 414 | 450 |

$$OR = \frac{Odds(D^+|E^+)}{Odds(D^+|E^-)} = \frac{\frac{P(D^+|E^+)}{P(D^-|E^+)}}{\frac{P(D^+|E^-)}{P(D^-|E^-)}} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

estimated by

$$\hat{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c} = \frac{30 \cdot 144}{6 \cdot 270} = 2.67$$

95% confidence interval

$$\hat{OR} \cdot \exp\left(-1.96 \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right) \text{ to } \hat{OR} \cdot \exp\left(1.96 \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

1.09 to 6.57

# Logistic regression model

- Logistic regression model in terms of ...

  ... conditional expectations (risiks):

$$P(D^+|x_1, x_2) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

... log-odds

$$\log(Odds(D^+|x_1, x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

# Interpretation

- If $x_1 = 0$ and $x_2 = 0$

$$\log(Odds(D^+|X_1 = 0, X_2 = 0)) = \beta_0$$

$$Odds(D^+|X_1 = 0, X_2 = 0) = \exp(\beta_0)$$

e.g. "chance for disease occurrence if there is no exposition to the risk factor or confounder"

- $$\log\left(\frac{Odds(D^+|E^+)}{Odds(D^+|E^-)}\right) = \log\left(\frac{Odds(D^+|X_1 = 1, X_2 = x_2)}{Odds(D^+|X_1 = 0, X_2 = x_2)}\right)$$

$$= \log(Odds(D^+|X_1 = 1, X_2 = x_2)) - \log(Odds(D^+|X_1 = 0, X_2 = x_2))$$

$$= \beta_0 + \beta_1 \cdot 1 + \beta_2 x_2 - (\beta_0 + \beta_1 \cdot 0 + \beta_2 x_2) = \beta_1$$

(or $OR = \exp(\beta_1)$ given exposition to the risik factor, taking the confounder into account)

# Cox proportional hazards regression

- Regression model for (censored) time to event data
- Considers the hazard (instantaneous risk) for an event at time $t$ given that no event has occurred to far

$$\lambda(t) = \lim_{h \to 0} \frac{P(t < T \leq t + h | T > t)}{h}$$

- Assumption:
  For two groups of patients A and B the hazard rates are proportional to each other:

$$\frac{\lambda_B(t)}{\lambda_A(t)} = const = RR = HR$$ hazard Ratio

  "proportional hazards model"

# Cox regression: model equation

- Regression model for hazard $\lambda$(t)

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 x_1 + \beta_2 x_2)$$

  $\lambda_0(t)$: baseline hazard

- Interpretation:
  exp($\beta_1$) = relative risk $x_1 = 1$ vs. $x_1=0$ ($x_2$ fixed)
  exp($\beta_2$) = relative risk $x_2 = 1$ vs. $x_2=0$ ($x_1$ fixed)

# More generally: risk prediction models

- Generalized linear models:
  - Observations $(x_i, y_i)$, i=1,…,$n$, with $x_i = (x_{i1},\ldots,x_{ip})'$

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

- Cox proportional hazards model
  - Observations $(t_i, \delta_i\varepsilon_i, x_i)$, i=1,…,$n$

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

- Estimation of parameter vector $\beta$ via (partial) log-likelihood $l(\beta)$

# Duality p-value – confidence interval

- A statistical test at significance level α can also be performedby checking whether the value of a statistical quantity assumed under the null hypothesis is contained in the (1-α)-confidence interval

# Multiple testing and family-wise error rate

- When performing $M$ tests at level $\alpha$, $M \cdot \alpha$ can be expected to be significant, even if there is no effect
- With a large number of tests, this is a severe problem, e.g., 400 genes would be expected to be significant in the Rosenwald data, even if there is no effect, when employing level $\alpha = 0.05$
- For ensuring an overall, i.e., family-wise, error rate $\alpha$, the Bonferroni method can be employed:
  Each test is performed at level $\alpha/M$
- **Example:** In the Rosenwald data, only two genes are significant at the level $0.05/7399 = 6.76 \cdot 10^{-6}$
- There are less conservative methods available, but with a large number of covariates, control of the family-wise error rate often is not what is actually wanted

# False discovery rate

- Possible outcomes of M hypothesis tests:

| | Called "not significant" | Called "significant" | Total |
|---|---|---|---|
| **H0 true** | U | V | M0 |
| **H0 false** | T | S | M1 |
| **Total** | M-R | R | M |

- False discovery rate: FDR = E(V/R)
- Controlling the false discovery rate instead of the family-wise error rate means that we accept that there will be a certain proportion of false positives, e.g., genes called significant that have no effect, and we attempt to control this proportion

# Approach of Benjamini and Hochberg

- Fix the false discovery rate α (e.g., α=0.15) and let $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M)}$ denote the ordered *p*-values
- Define

$$L = \max\{j: p_{(j)} < \alpha \cdot j/M\}$$

- Reject all hypotheses for which $p_j \leq p_{(L)}$
- If the hypotheses are independent, Benjamini and Hochberg (1995) show that

$$FDR \leq M_0/M \cdot \alpha \leq \alpha$$

- Later work …
  - relaxes the assumption of "independent hypotheses"
  - provides tighter control of the FDR by employing estimates of $M_0$

# Stepwise variable selection

- Various schemes for modifying a model one at a time (add or remove covariate), based on *p*-values:
  - forward selection
  - backward elimination
  - univariate screening + backward elimination
- Recommendations highly subjective

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

# Stepwise and stagewise variable selection

| | stepwise | stagewise |
|---|---|---|
| **approach** | focussing one covariate in each model building step | |
| **in each step** | add/remove one covariate | update the parameter for one covariate |
| **increase of complexity** | 1 degree of freedom | < 1 degree of freedom |
| **other covariates** | re-estimate parameters for all included | keep fixed |
| **stopping rule** | p-value or prediction performance | prediction performance |
| **tuning parameter** | (stopping rule) | number and size of steps |

# Bias-variance tradeoff

- Why should one want biased estimates?
  - Smaller prediction error
  - Closely connected: more precise parameter estimates
- Bias in parameter estimation should depend on true parameter values
  - For small true values bias can be large, because these cannot be estimated well anyway
  - For large true values there should be no/hardly any bias

# Misclassification rate and ROC curves

- When a risk prediction models provides a 0/1 classification, the performance can be evaluated via the misclassification rate

$$\frac{1}{n}\sum_{i=1}^{n} I(\hat{y}(x_i) \neq y_i)$$

  with respect to the true response
- Often, a continuous prediction will be provided. Applying different cutoffs for classification, different values are obtained for the
  - **sensitivity**, i.e., the proportion of true 1s that are classified as 1
  - **specificity**, i.e., the proportion of true 0s that are classified as 0
- Tracking these results in an ROC curve, a larger area under this curve (AUC) indicates better performance

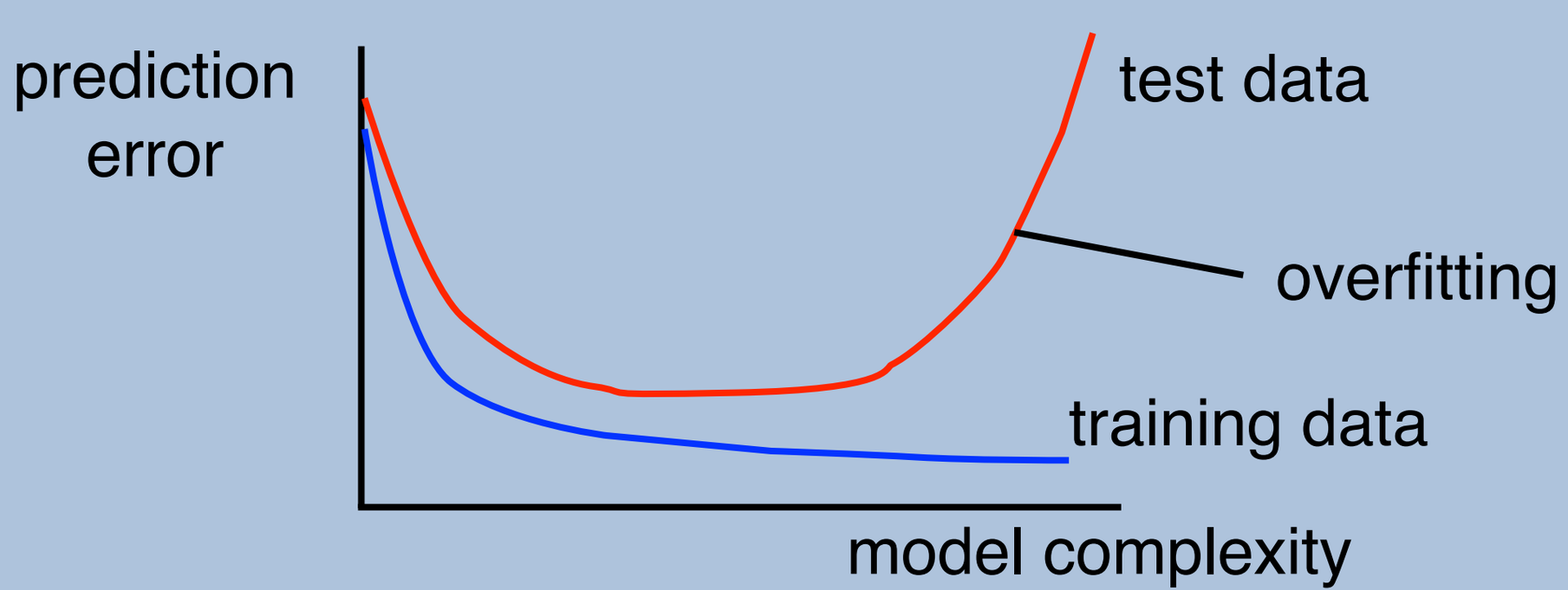
# The Brier score

- Often, not only a continuous prediction will be provided, but a predicted probability $P(Y = 1|x)$ via a fitted risk prediction rule $\hat{r}(x_i)$
- The Brier score then is $E[(Y - \hat{r}(x))^2]$
- In contrast to the misclassification rate, ROC curves and AUC it is a **strictly proper scoring rule**, i.e., it takes its minimum value only if the predicted probabilities are equal to the true probabilities
- The empirical Brier score is calculated as

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{r}(x_i))^2$$

# Selecting model complexity

- Often techniques for fitting risk prediction models have a tuning parameter that determines model complexity
- Ideally these should be chosen to maximize prediction performance on new test data:




- Problem: Test data will often not be available. All data should be used for model fitting!

# Cross-validation

- For getting a stable selection of model complexity:
  - Repeatedly split the data into training and test observations
  - Fit the risk prediction model for various levels of complexity to the training observations
  - Evaluate prediction performance using the test observations
  - Average the performance estimates for each level of complexity over all splits
  - Chose that level of complexity with maximal performance
- In traditional *k*-fold cross-validation, the data is split into *k* folds. Each fold serves once as a test set, while the remaining folds constitute the training set

# Resampling for model evaluation

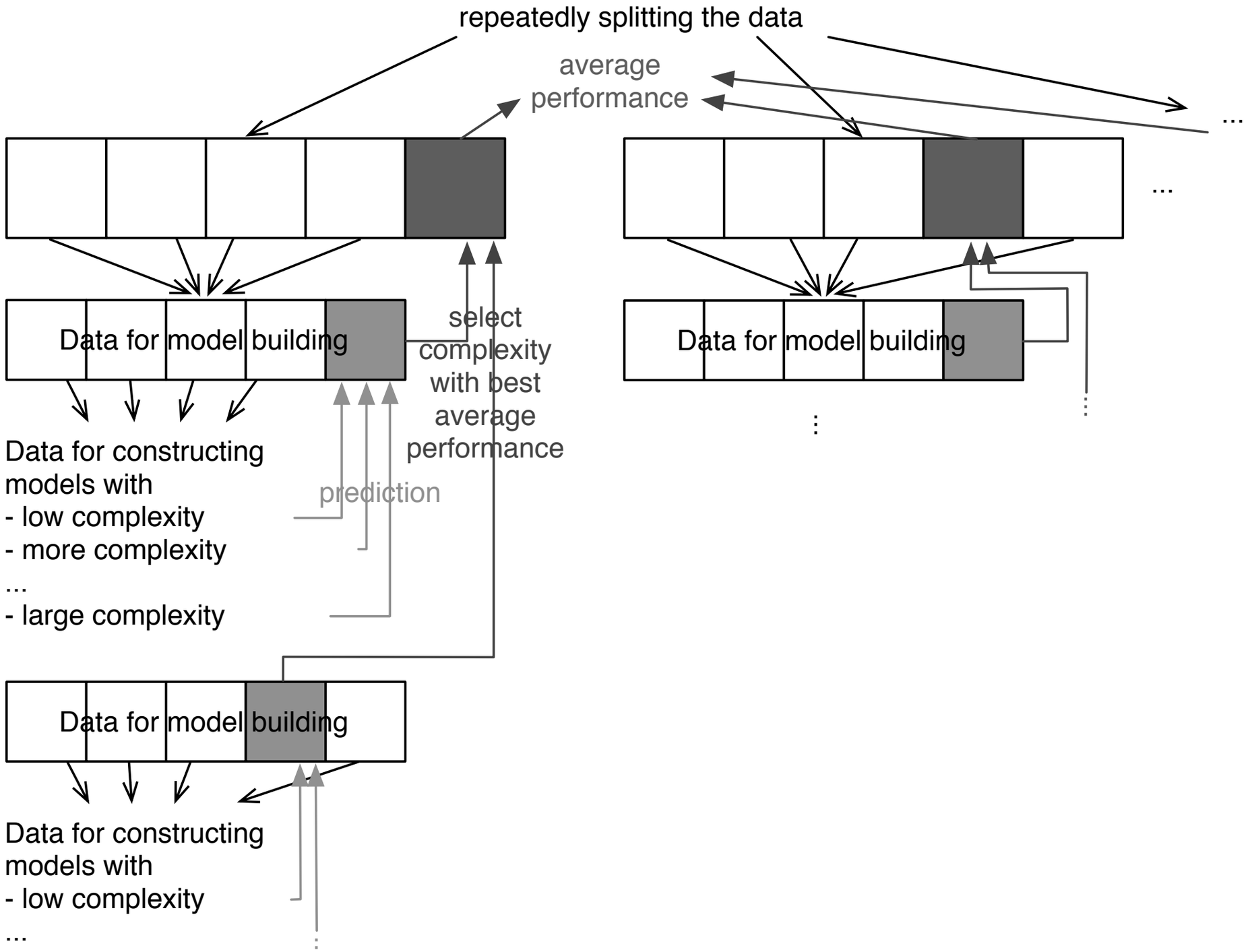
When the number of covariates is large, "perfect" prediction can always be obtained on the data that was used for fitting a risk prediction model
→ not useful for judging prediction performance in new data

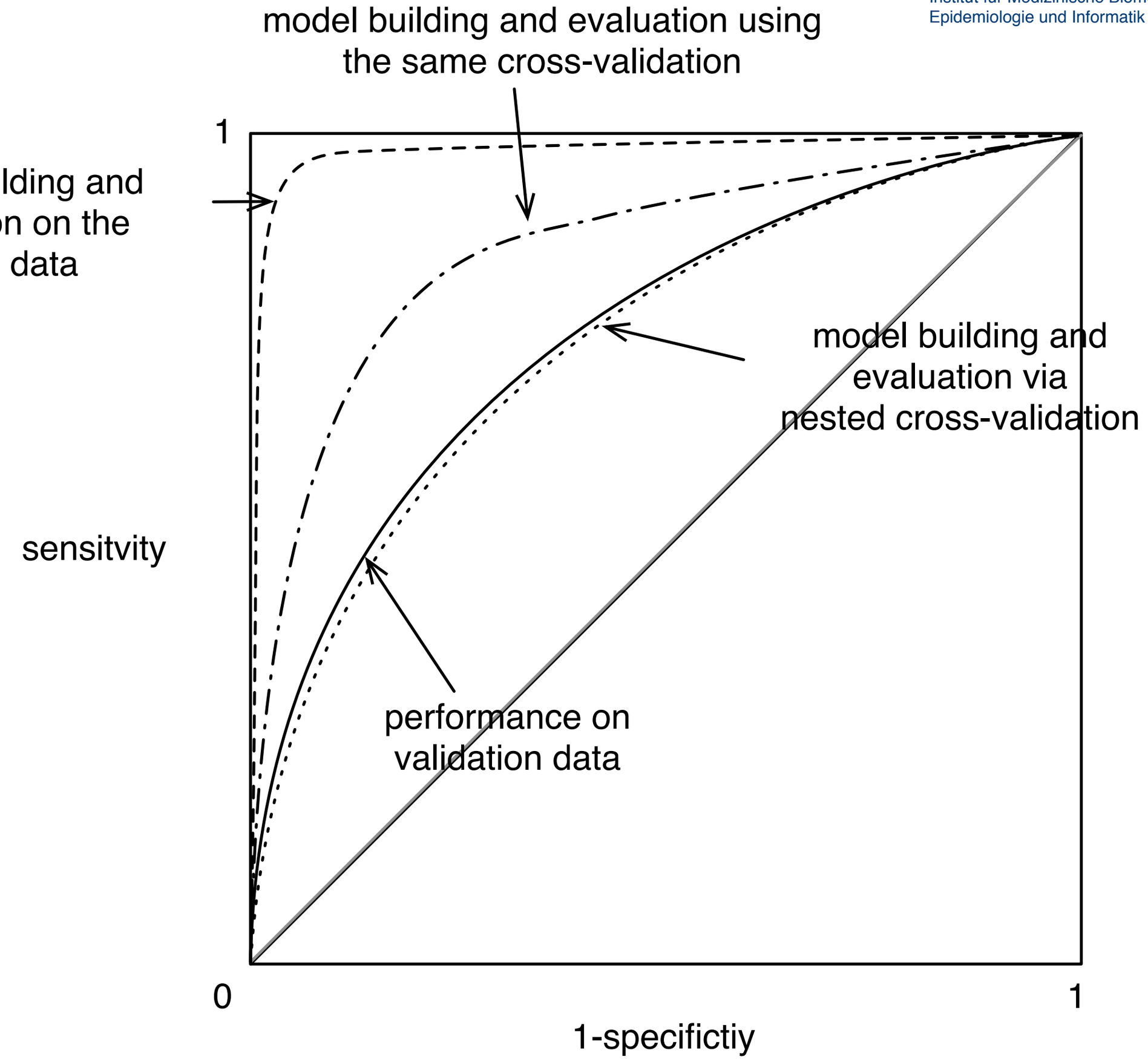
Alternatives:

- Set aside test data → too expensive
- Resampling (bootstrap, cross-validation): Repeatedly, generate "new" data by randomly drawing observations, and evaluate on left-out observations
  Important: all model building steps in each resampling data set

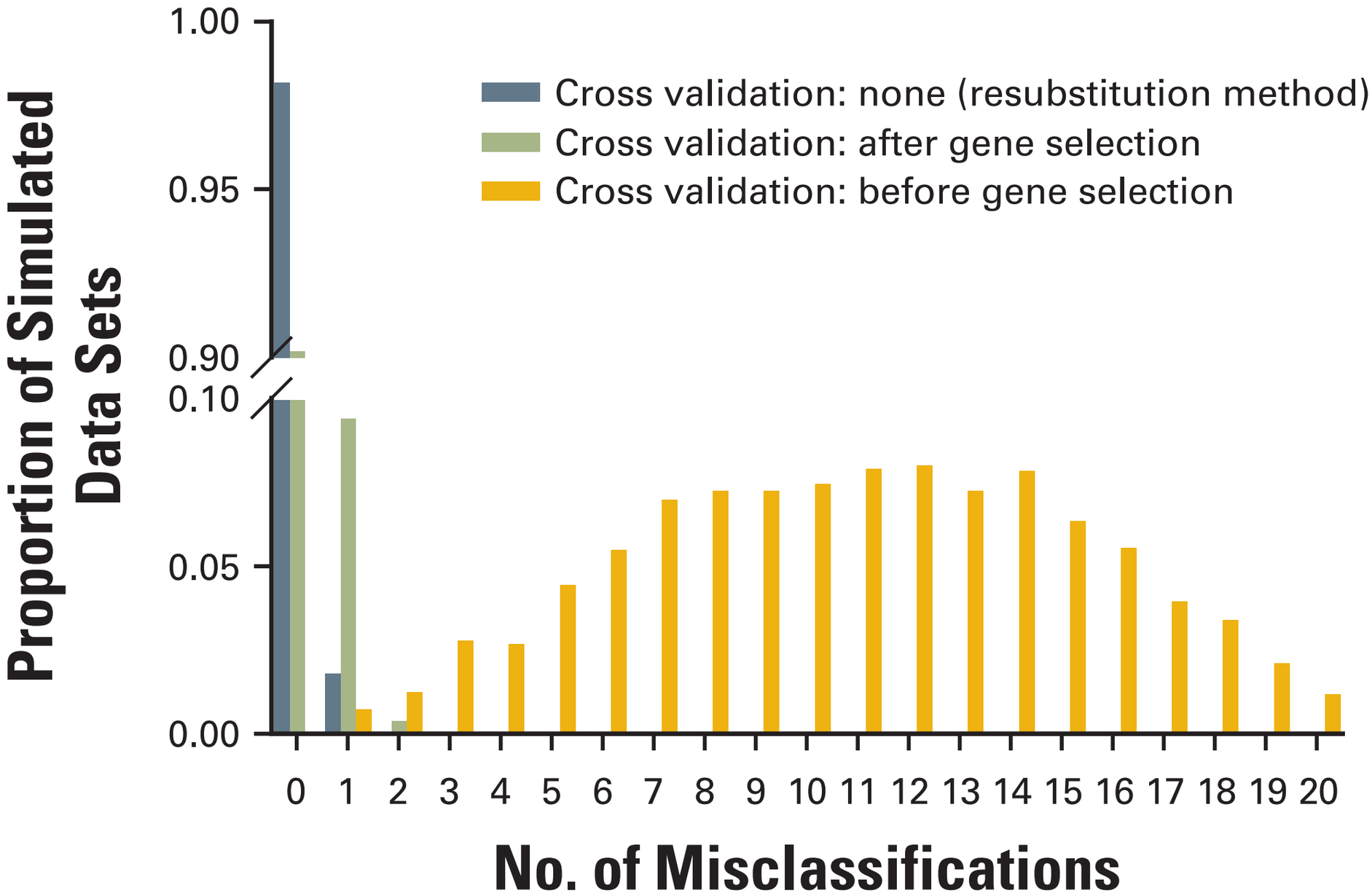
UNIVERSITÄTSmedizin. MAINZ
Institut für Medizinische Biometrie,
Epidemiologie und Informatik

# Nested cross-validation




model building and evaluation using the same cross-validation
model building and evaluation on the same data
model building and evaluation via nested cross-validation
performance on validation data
1
0
1
sensitvity
1-specifictiy

# Simulation study by Simon et al. (2003)



# R demo

- Start R