MEDICAL
UNIVERSITY
OF VIENNA

GENE SELECTION IN MICROARRAY SURVIVAL STUDIES UNDER POSSIBLY NON-PROPORTIONAL HAZARDS

Daniela Dunkler, Michael Schemper and Georg Heinze

Section for Clinical Biometrics
Center for Medical Statistics, Informatics and Intelligent Systems
Medical University of Vienna, Austria

Project sponsored by the Austrian Research Fund FWF

***Bioinformatics* 26(6); 784-790, 2010**

Our Motivation I

- Given: high-dimensional gene expression data with survival outcome (like Rosenwald *et al.* N Engl J Med, 2002)

- Goal: identify genes possibly linked to survival

- Talk: limited to univariate gene selection, but methods generalize to other gene selection methods.

Our Motivation II

- Typical analysis: **Cox regression**
- Cox regression assumes **proportional hazards**:

  = A constant effect of gene expression on survival over the whole period of follow-up.

- Problem: Proportional hazards assumption may be questionable, but cannot be verified for all genes.

- Ignoring the proportional hazards assumption:
  - **Cox regression will lead to over- and underestimation for a considerably number of genes.**
  - Cox regression hazard ratios are not directly comparable.

A possible Solution

We need a *summary measure of effect size* which is suitable to rank genes when some genes may exhibit a time-dependent effect on survival.

➡ ***generalized concordance probability***

Outline

- Concordance probability c
- Generalized concordance probability c' for continuous data
- Two methods to estimate c'
  - Concordance regression
  - Weighted Cox regression
- Comparison of Cox, concordance and weighted Cox regression
  - in Monte Carlo Study
  - analyses of real data
- Extensions
- Conclusions

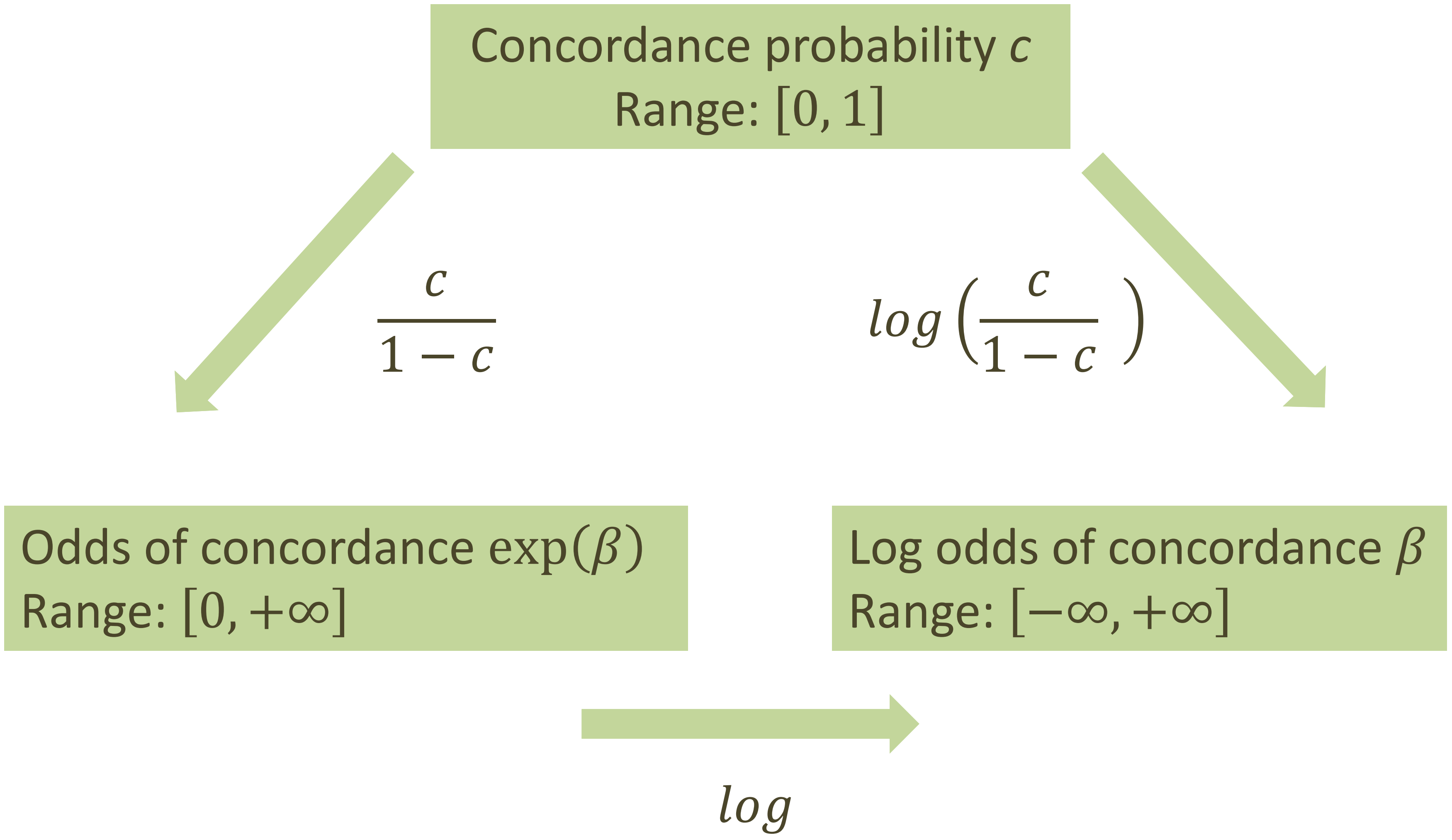
Concordance probability c

- Consider 2 groups:
- c = non-parametric measure of separation of the survival distributions:
$$c = P(T_1 < T_0)$$
- Uncensored data: $c \equiv$ Mann-Whitney statistic

- Under proportional hazards:
  - Cox regression hazard ratio = $\exp(\beta) = c/(1-c)$
- Under non-proportional hazards:
  - $\exp(\beta) \neq c/(1-c)$
  - c still has an intuitive interpretation

Odds of concordance

Concordance probability c



Concordance probability c
Range: $[0, 1]$

$\frac{c}{1-c}$

$log\left(\frac{c}{1-c}\right)$

Odds of concordance $\exp(\beta)$
Range: $[0, +\infty]$

Log odds of concordance β
Range: $[-\infty, +\infty]$

log

Generalized concordance probability c'

- Consider a continuous variable X:
- Define $\Gamma(x_i, x_j) = \text{logit}\left[P\{T(x_i) < T(x_j)\} \right]$ as the log odds of concordance between two individuals with arbitrary log-2 gene expression values x_i and x_j.
- Assume that $\Gamma(x_i, x_j) \propto (x_i - x_j)$ $\triangleq$ **Linearity assumption**
- Implies $\Gamma(x_i, x_j) / (x_i - x_j) = \gamma$ irrespective of the actual values of x_i and x_j.
- The generalized concordance probability c' is

$$c' = \frac{\exp(\gamma)}{1+\exp(\gamma)} = P\{T(X = x+1) < T(X = x)\}$$

Concordance regression I

- Model c' by conditional logistic-type (*concordance*) regression:

$$P\left[T(x_i) < T(x_j)\right] = \frac{\exp(x_i\beta)}{\exp(x_i\beta) + \exp(x_j\beta)}$$

- The derivative of the conditional logistic log likelihood:

$$\partial\ell / \partial\beta = \sum_{(i,j)} [x_i - \frac{x_i \exp(x_i\beta) + x_j \exp(x_j\beta)}{\exp(x_i\beta) + \exp(x_j\beta)}],$$

- Summation: over all available 'risk pairs' (i, j) such that $t_i < t_j$.

- β denotes the $\text{logit}\left[P\{T(x_i) < T(x_j)\}\right]$ related to a one-unit increase in X

➡ $\hat{\beta}$ directly estimates $\hat{\gamma}$

➡ $\hat{c}' = \exp(\hat{\beta}) / \{1 + \exp(\hat{\beta})\}$

Concordance regression II

- No censoring:
  - Each individual appears in n-1 ‘risk pairs ’.

- Censoring:
  - Omit all risk pairs where the shorter time t_i is censored

    ➡ Overrepresentation of some individuals

    ➡ Weight the remaining risk pairs by their inverse sampling probabilities.

Concordance regression III

- Weight function: Assume $t_i < t_j$

# of risk pairs with subject *i* dying earlier
had censoring not occured

$$w(i,j) = \frac{N(0)S(t_i) - 1}{N(t_i) - 1} \times G(t_i)^{-1}$$

# of risk pairs with
subject *i* dying earlier

Compensates the attenuation
in observed events due to
earlier censorship

$\mathrm{N}(t)$= # of subjects at risk at time t
$\mathrm{S}(t)$ = left continuous Kaplan Meier estimate at time t
$\mathrm{G}(t)$= Kaplan meier estimate with the status indicator reversed at time t

Weighted Cox regression I

- Schemper *et al.* (Stat. Med 2009) introduce weights into the score function to obtain average hazard ratio = $\exp(\beta)$
- The weights are chosen to maintain the interpretability of estimates under non-proportional hazards:
- Over a wide range of β: $\exp(\beta) \sim \exp(\gamma)$

Weighted Cox regression II

- The weights are defined by

$$w(t_i) = S(t_i) \times G(t_i)^{-1}$$

Reflects the relative importance attributed to the log hazard ratio at time *t*

Compensates the attenuation in observed events due to earlier censorship

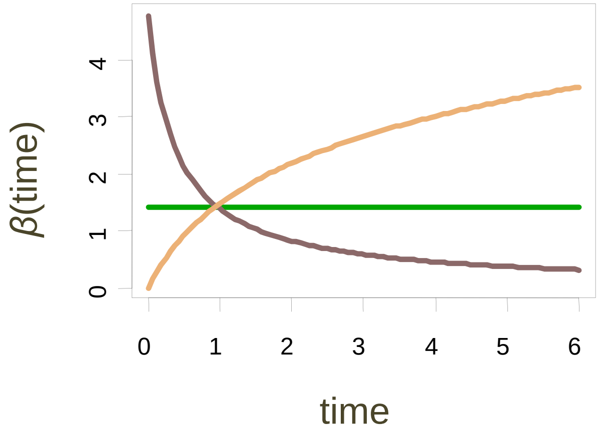
S(*t*) = left continuous Kaplan Meier estimate at time *t*

G(*t*) = Kaplan meier estimate with the status indicator reversed at time *t*

‘Univariate’ Simulation

- Match gene expression [N(0, 1)] to marginal failure times [Weibull(2, 0.5)] by algorithm of MacKenzie and Abrahamowicz (Stat Comput, 2002)

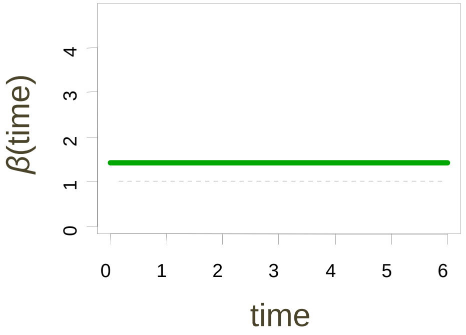
- Type of time-dependency
  - Proportional hazards
  - Diverging hazards
  - Converging hazards



- Varied amount of censoring and effect sizes
- 2000 samples of 200 observations

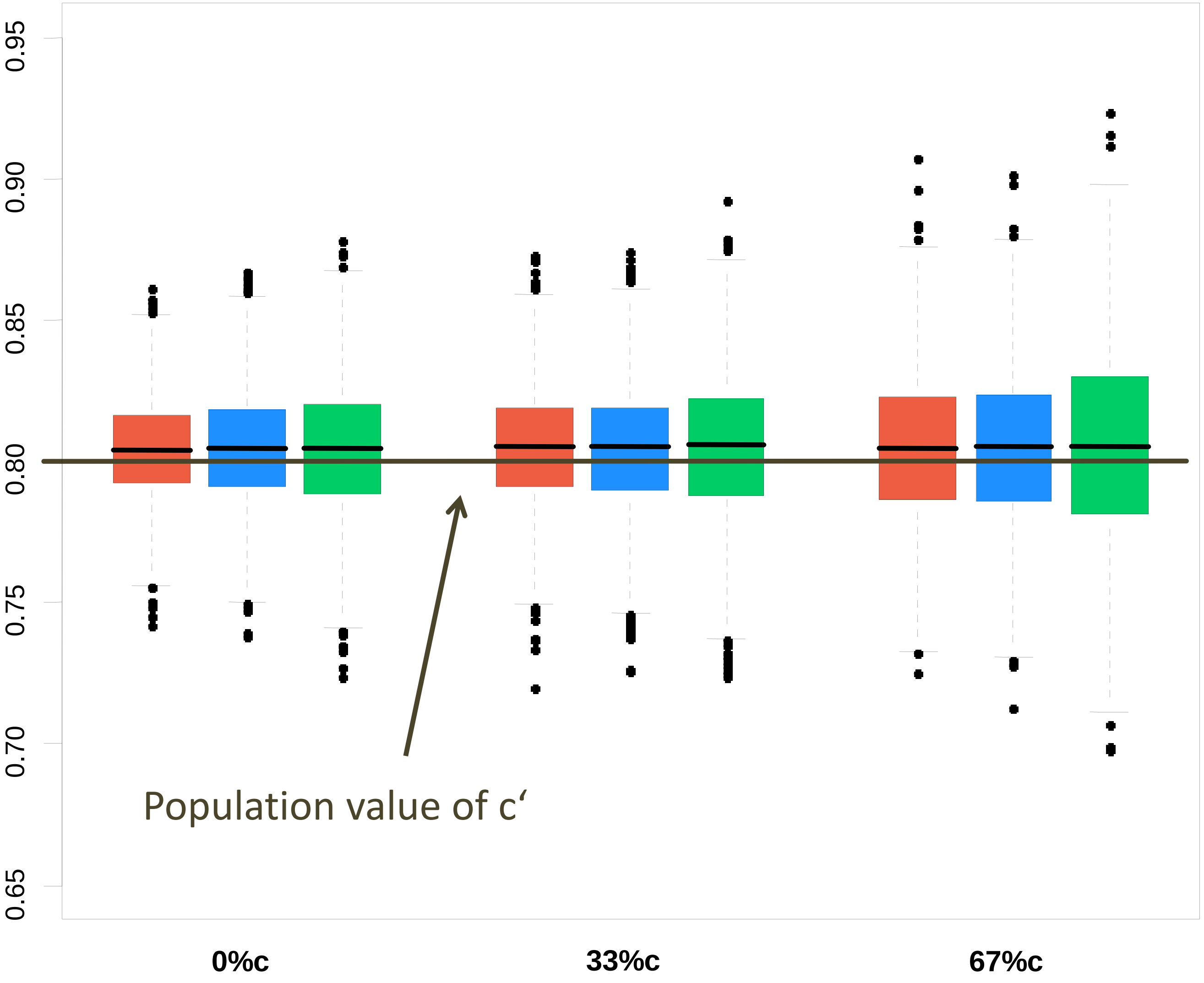
- For each sample and each method univariate models are fit.

Proportional hazards



Effect size:

$$c' = 0.8 \triangleq$$
$$\triangleq \beta = \log(4)$$



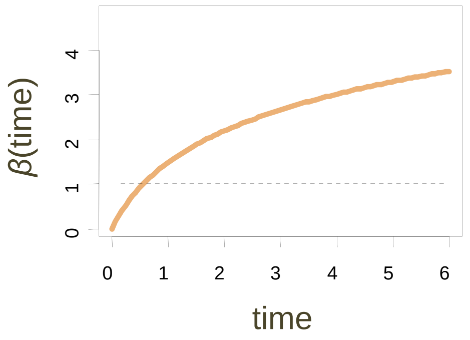
Population value of c‘

Cox regression

Weighted Cox reg.

Concordance reg.

Diverging hazards



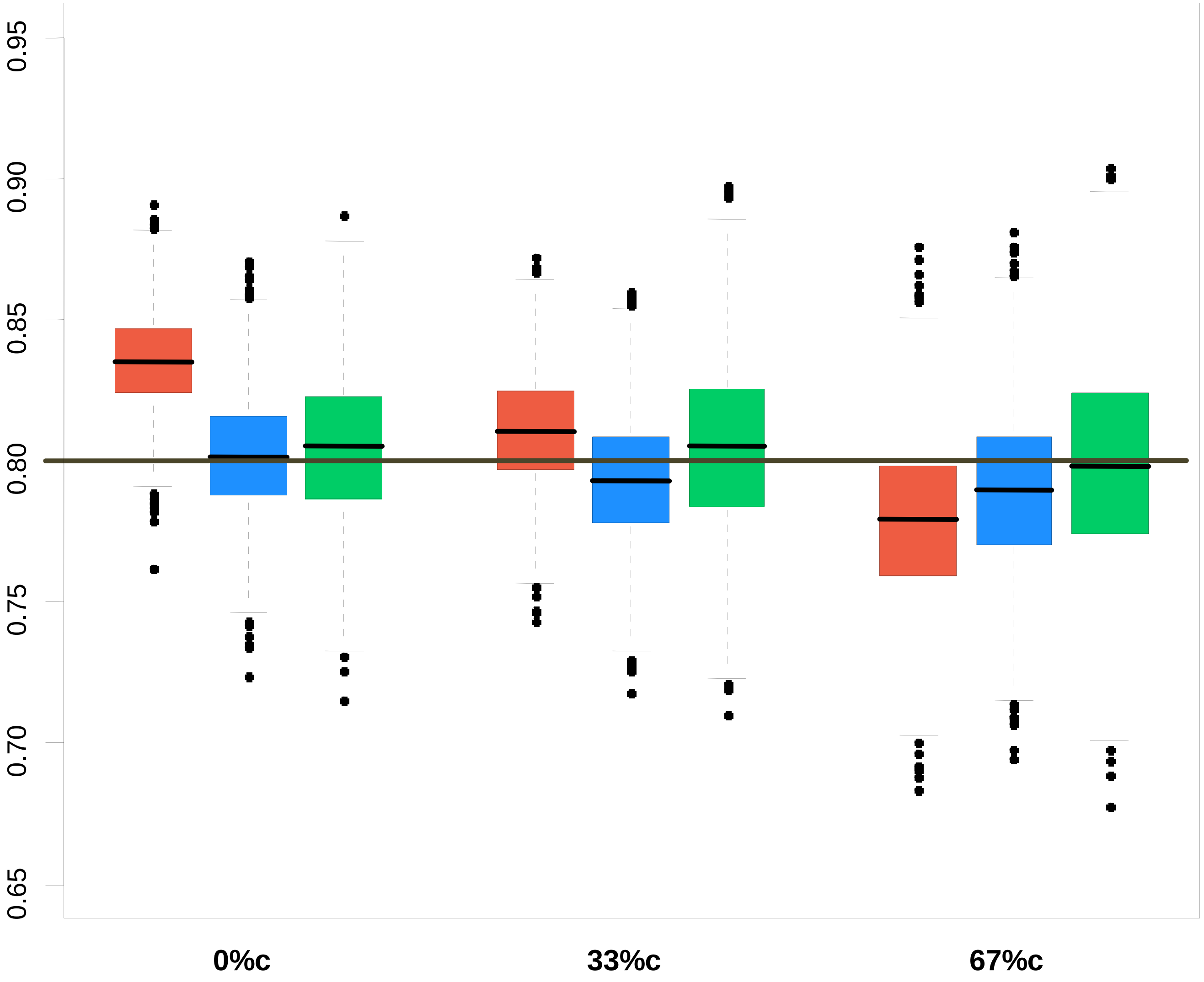
Effect size:

$$c' = 0.8$$

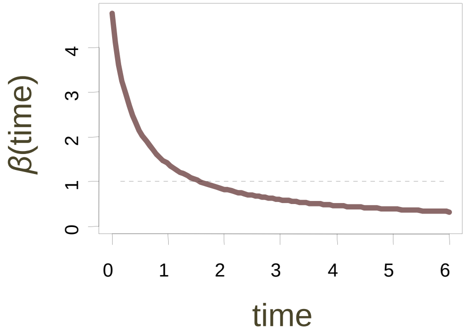
Cox regression

Weighted Cox reg.

Concordance reg.

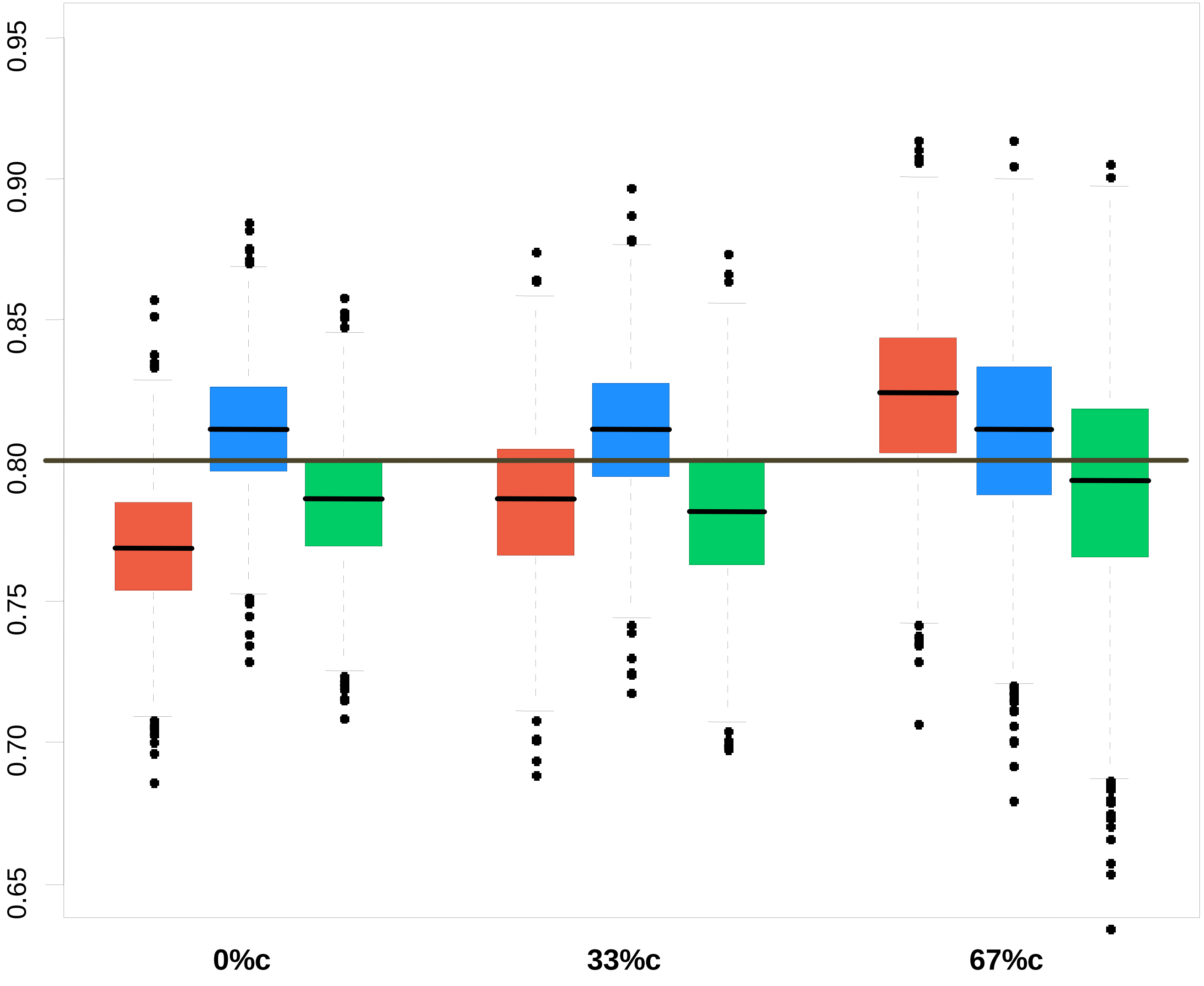


Converging hazards



Effect size:

$$c' = 0.8$$



Cox regression

Weighted Cox reg.

Concordance reg.

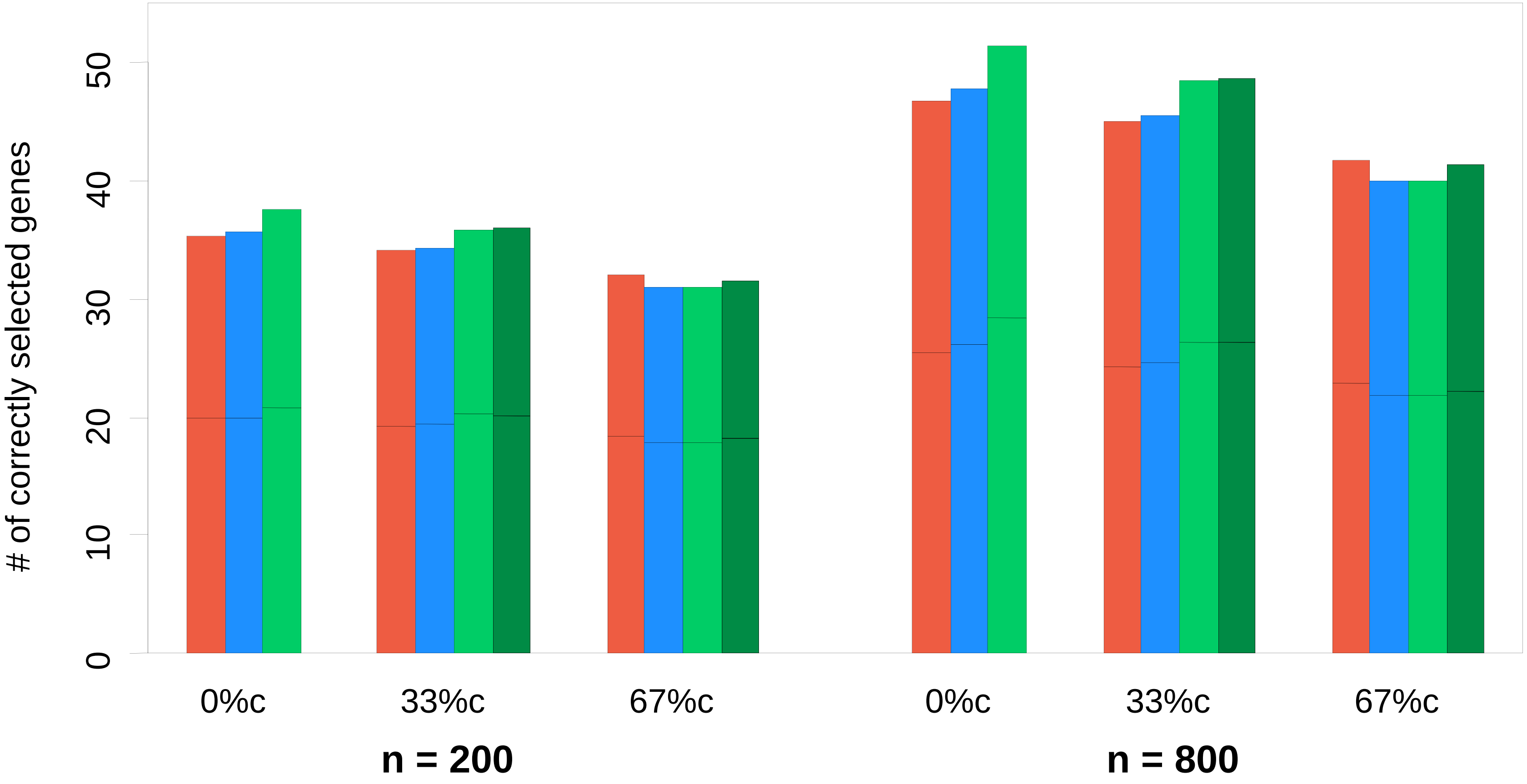
‘Multivariate’ Simulation

- **Mimic real-life gene expression data:**
  - according to Binder and Schumacher (Stat Appl Genet Mol Biol, 2008)
  - 72 of 5000 genes have additive effect on log hazard:
    - 1/3 with proportional hazards
    - 1/3 with diverging hazards
    - 1/3 with converging hazards
  - Varied amount of censoring and sample size

1) Rank genes by univariate absolute effect size.
2) ‘Select’ 72 top genes for each method.
3) Compare the true positive rates.

‘Multivariate’ Simulation II

Select 72 genes from 5000 candidate genes



Cox regression

Weighted Cox reg.

Concordance reg.

Concordance reg. + truncation of weights

‘Multivariate’ Simulation

- **Mimic real-life gene expression data:**
  Gene selection should depend on effect size,
  not on type of time-dependency and/or censoring:

  + **Concordance regression**

  ~ **Weighted Cox regression:** prefers converging hazards

  ~ **Cox regression:** dependent on censoring

Application to real-life data I

Bhattacharjee *et al.* data
(PNAS, 2001)

- Lung adenocarcinomas
- Patients: 125
- Survival endpoint: 71
- Genes: 12600

Rosenwald *et al.* data
(N Engl J Med, 2002)

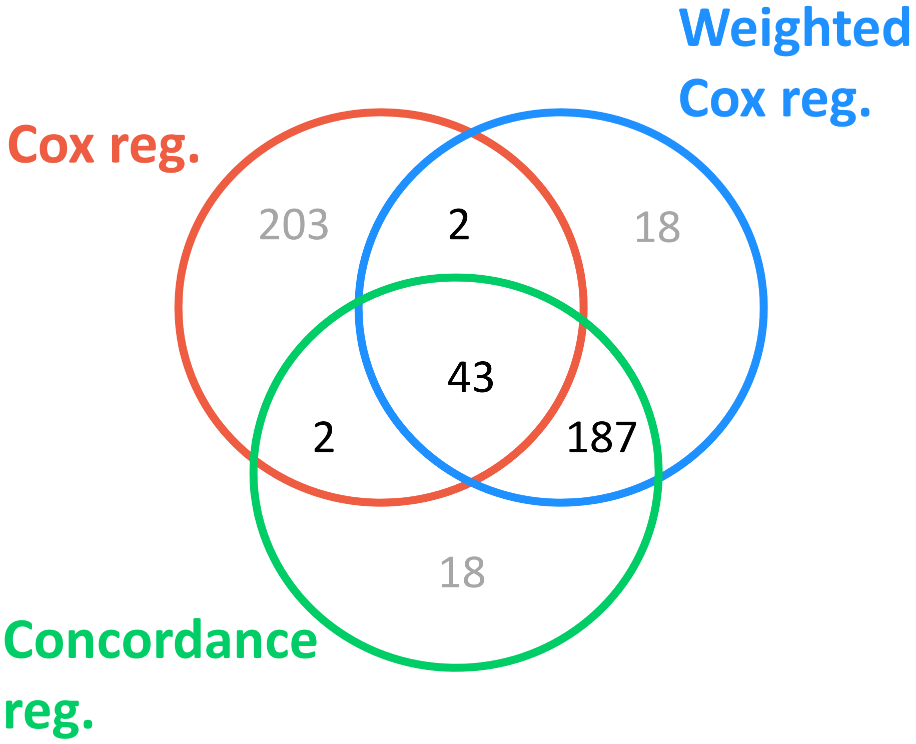
- Diffuse large B-cell lymphoma
- Patients: 240
- Survival endpoint: 138
- Genes: 7053

1) For each gene and each method fit univariate models.
2) Rank genes by absolute effect size.
3) ‘Select’ the 250 top genes for each method.

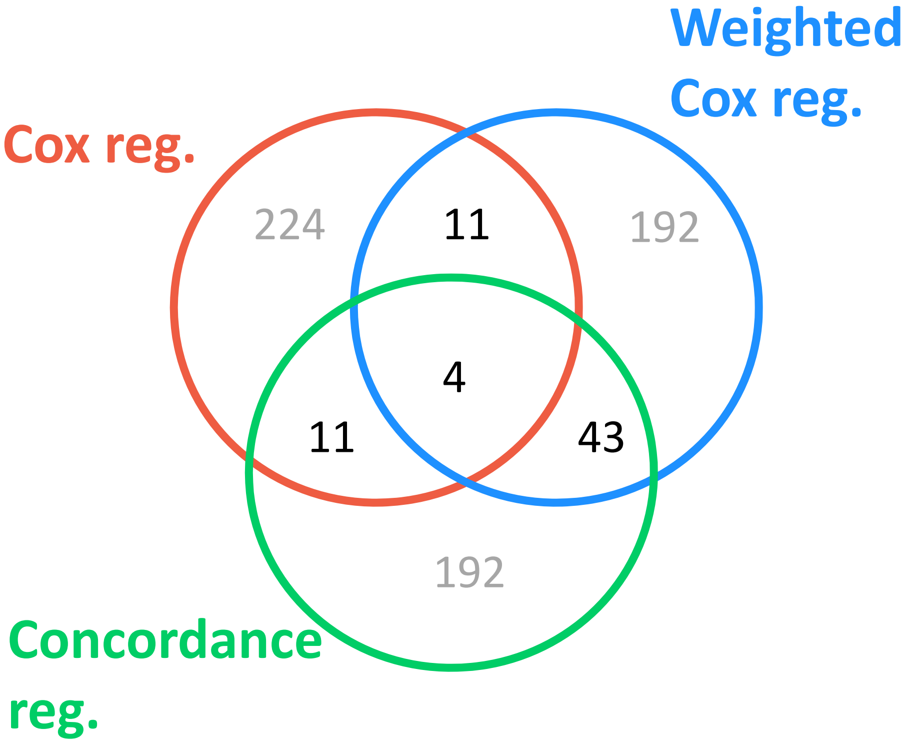
Application to real-life data II

‘Select‘ 250 top genes …

Bhattacharjee *et al.* data



Rosenwald *et al.* data



Extensions: multivariable modeling with concordance regression

- So far only univariate modeling was discussed
- Multivariable models straightforward
- Regularization (LASSO, ridge, elastic net) possible via *penalized* R package: selection and prediction
  Regularized concordance regression
  - may provide more robust models than regularized Cox regression
  - is less dependent on censoring pattern, more generalizable to other validation cohorts or populations
  - can be used for sensitivity analysis
  - or for enrichment of a gene set found by regularized Cox regression

Extensions: nonparametric *c*

- Semi-parametric: $c' = P\left(T_i < T_j | X_i = X_j + 1\right)$
- Non-parametric: $c = P\left(T_i < T_j | X_i > X_j\right)$
  - Harrell (1982)
  - Assessing relationship of a prognostic index with survival
  - Applied in Ma & Xiao (Brief Bioinform, 2010)
  - Robust to misspecifications

Conclusions

- We propose to use c' as a summary measure of effect size to rank genes irrespective of the type of time-dependency and censoring pattern.
- c' is a concise single number useful for clear decisions at time 0.

- **Concordance regression** gives the least biased and most stable estimates irrespective of type of time-dependency and censoring pattern.

- Software implementation: R packages
  - Weighted Cox regression: **`coxphw`** (available at CRAN)
  - Concordance regression: **`concreg`** (semiparametric c' and nonparametric c; available at CRAN)