Supplementary Data

Gene selection in microarray survival studies under possibly non-proportional hazards

Daniela Dunkler, Michael Schemper and Georg Heinze

January 2010

Daniela Dunkler
Section of Clinical Biometrics
Center for Medical Statistics, Informatics and Intelligent Systems
Medical University of Vienna
Spitalgasse 23, A-1090 Vienna, Austria
email: daniela.dunkler@meduniwien.ac.at
phone: +43-1-40400-6682

Abstract

Motivation: Univariate Cox regression (COX) is often used to select genes possibly linked to survival. With non-proportional hazards (NPH), COX could lead to under- or overestimation of effects.

The effect size measure $c = P(T_1 < T_0)$, i. e. the probability that a person randomly chosen from group G_1 dies earlier than a person from G_0, is independent of the proportional hazards (PH) assumption. Here we consider its generalization to continuous data c' and investigate the suitability of c' for gene selection in microarray survival studies.

Results: Under PH, c' is most efficiently estimated by COX. Under NPH, c' can be obtained by weighted Cox regression (WHE) or a novel method, concordance regression (CON). The least biased and most stable estimates were obtained by CON. We propose to use c' as summary measure of effect size to rank genes irrespective of different types of NPH and censoring patterns.

Availability: WHE and CON are available as R packages.

Contact: daniela.dunkler@meduniwien.ac.at

Contents

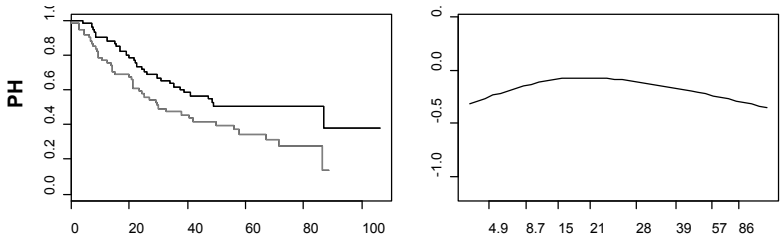
Abstract 2
Contents 3
1. Introduction 5
2. Exemplification of genes with proportional, converging or diverging hazards 5
3. Real-life applications 6
4. Simulation study 13
4.1. Univariate evaluation 13
4.2. Multivariate evaluation 17
Multivariate Evaluation No. 1 21
Multivariate Evaluation No. 2 23
Multivariate Evaluation No. 3 25
Multivariate Evaluation No. 4 27
Multivariate Evaluation No. 5 28
Multivariate Evaluation No. 6 29
p-values from paired t-tests for method comparisons corresponding to Table 1 30
5. Abbreviations 31
6. References 32
Appendix: Data generating algorithm 33

1. Introduction

The Supplementary Data to '*Gene selection in microarray survival studies under possibly non-proportional hazards*' gives additional results of the analysis of three real-life microarray data sets and the simulation study.

2. Exemplification of genes with proportional, converging or diverging hazards

Web-Figure 1 shows three examples of genes with proportional hazards (PH), converging hazards (CH) and diverging hazards (DH) from the study of Bhattacharjee *et al.* (2001). For the gene with PH (Web-Figure 1 row 1) the correlation of the scaled Schoenfeld residuals with the rank of time is close to 0. For the other genes the correlation is considerably larger indicating violation of the PH assumption. Web-Figure 1 row 2 shows a gene with CH, where the effect fades out with time, whereas the gene depicted in Web-Figure 1 row 3 exhibits DH, i. e., an effect increasing with time. Note, Cox regression (COX) results are only valid in case of PH.



3. Real-life applications

We applied univariate Cox regression (COX), weighted Cox regression (WHE) and concordance regression (CON) to all genes of three well-known freely-available microarray data sets and evaluated differences in gene selection. The data sets investigated were:

(i) Beer *et al.* (2002) studied the association of survival and gene expression profiles of microarray data of patients with early stage lung adenocarcinomas. The data were downloaded from http://dot.ped.med.umich.edu:2000/ourimage/pub/Lung/index.html (access date 18 January 2008)*. Negative gene expressions were set to 0.1. All gene expression values were $\log_2$-transformed and standardized. Only the 4966 genes used in the original publication were selected for further analysis. Survival times were available for 86 patients (24 events).

(ii) Similarly, Bhattacharjee *et al.* (2001) investigated correlation of gene expression from lung adenocarcinomas with a survival endpoint. The data ('DatasetA_12600gene.xls') were downloaded from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC61120/bin/pnas_191502998_index.html (link available at 14 January 2010). This data set had already been pre-processed and normalized. For a detailed description of the preprocessing we refer to http://www.ncbi.nlm.nih.gov/pmc/articles/PMC61120/bin/pnas_191502998_1.html (link available at 14 January 2010). Gene expression values were standardized to unit standard deviation. We used all 12600 gene expressions available for 125 patients (71 events) with survival information in our analysis. Ties in event times were arbitrarily broken.

(iii) In a study by Rosenwald *et al.* (2002) the association of gene expression and survival in 240 patients with diffuse large B-cell lymphoma was investigated. The data were downloaded from http://llmpp.nih.gov/DLBCL/ (link available at 14 January 2010). The downloadable data had already been preprocessed and $\log_2$-transformed. Only genes with at least 60% non-missing values were used in the analysis. Missing values were imputed by the k-nearest neighbour method with k=10, then the gene expression values were standardized. In the analysis 7053 genes were used. Among the 240 patients 138 events were observed. Ties in event times were arbitrarily broken.

* In January 2010 this link no longer works. The data are available from the author upon request.

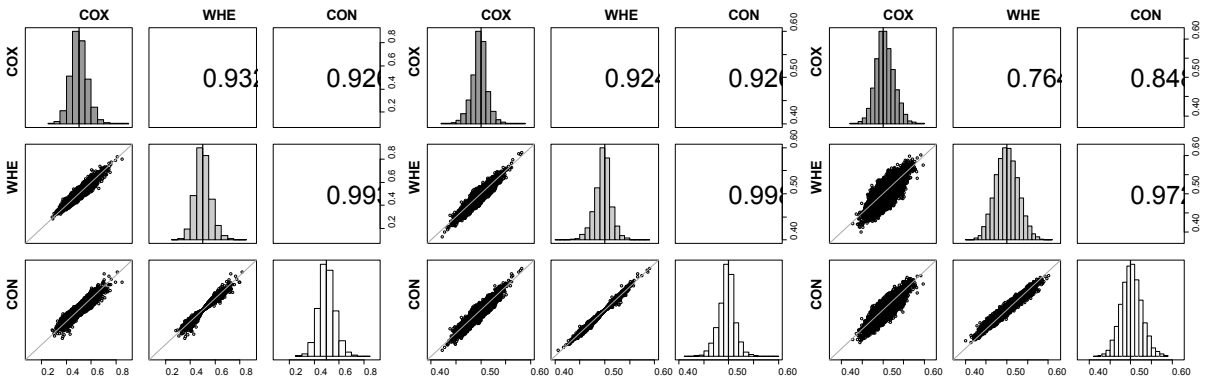
In each data set we ranked genes by their estimated absolute effect size, i. e., by $c'_+ = 0.5 + |c' - 0.5|$. We determined the threshold value $\hat{c}'_{+(250)}$ such that a predetermined number of 250 'selected' genes exceed this value in their absolute effect size. The number of false positive selections FP was estimated as the average number of selected genes (with $\hat{c}'_{+(250)}$ as threshold) in 100 versions of the data set that resulted from permuting the survival information. The proportion of genes not linked to survival was estimated as

$$\hat{\pi}_0 = \sum_{g=1}^{G} I\{\hat{c}'_g \in (q_{25}, q_{75})\} / (0.5G)$$

where q_{25} and q_{75} are the 25th and 75th percentiles of the permutation distribution of $\hat{c}'$ across all G genes and B permutations, and $\hat{c}'_g$ is the original data estimate of gene g. Web-Table 1 gives $\hat{\pi}_0$ for the three analyzed real-life data. The false discovery rate FDR_{250} was then calculated as $\mathrm{FDR}_{250} = \mathrm{FP} \times \hat{\pi}_0 / 250$.

	$\hat{\pi}_0$		
	COX	WHE	CON
Beer	0.83971	0.82642	0.96899
Bhattacharjee	0.98603	0.95127	0.96412
Rosenwald	0.82291	0.79711	0.87169

Web-Table 1: The proportion of genes not linked to survival $\hat{\pi}_0$ for three analyzed real-life data. COX, Cox regression; WHE, weighted Cox regression; CON, concordance regression.



Web-Figure 2: Comparison of estimated by COX, WHE and CON for three real-life data sets. The numbers represent the Spearman correlation coefficient. COX, Cox regression; WHE, weighted Cox regression; CON, concordance regression.

Web-Figure 2 shows that $\hat{c}'$ estimated with WHE and CON are highly correlated in all three data sets. The correlation with COX and the other two methods is always smaller.

Web-Table 2 summarizes the results of COX, WHE and CON for the real-life data. In all three data sets approximately half of the genes have a negative effect on survival. The range of $\hat{c}'$ varies considerably between the three data sets, with the largest range in the Beer data set and the smallest range in the Bhattacharjee data set. The correlation of the absolute effect size estimated with WHE and CON is close to 1 for all data sets, whereas the correlation between the absolute effect size estimated with COX and WHE or COX and CON is considerably smaller.

If 250 genes with the largest absolute effect size are selected with each method the largest agreement in gene selection is observed between WHE and CON. In all three data sets approximately 50% of the selected genes are selected by COX, WHE and CON. The FDR_{250} if 250 genes are selected exhibits no clear favourite method. For two data sets COX reaches the smallest FDR_{250}, but the ranking of COX may be biased because COX assumes proportional hazards, which cannot be proven for all genes. With some thresholds m the FDR_m was larger than 1, a situation which was already anticipated by Tusher *et al.* (2001).

			Data		
	Statistic	**Method**	**Beer**	**Bhattacharjee**	**Rosenwald**
General information	*# observations*		*86*	*125*	*240*
	# events		*24*	*71*	*138*
	# genes		*4966*	*12600*	*7053*
	# of genes with $\hat{c}' < 0.5$	COX	2434 (49%)	5360 (43%)	4044 (57%)
	# of genes with $\hat{c}' < 0.5$	WHE	2453 (49%)	5193 (41%)	3511 (50%)
	# of genes with $\hat{c}' < 0.5$	CON	2450 (49%)	5193 (41%)	3672 (52%)
	Range of $\hat{c}'$	*COX*	*0.283-0.850*	*0.423-0.583*	*0.426-0.598*
	Range of $\hat{c}'$	*WHE*	*0.283-0.819*	*0.406-0.590*	*0.401-0.601*
	Range of $\hat{c}'$	*CON*	*0.254-0.828*	*0.412-0.591*	*0.416-0.590*
	Cor of $\hat{c}'_+$ by	COX & WHE	0.792	0.829	0.464
	Cor of $\hat{c}'_+$ by	COX & CON	0.767	0.829	0.613
	Cor of $\hat{c}'_+$ by	WHE & CON	0.972	0.994	0.902

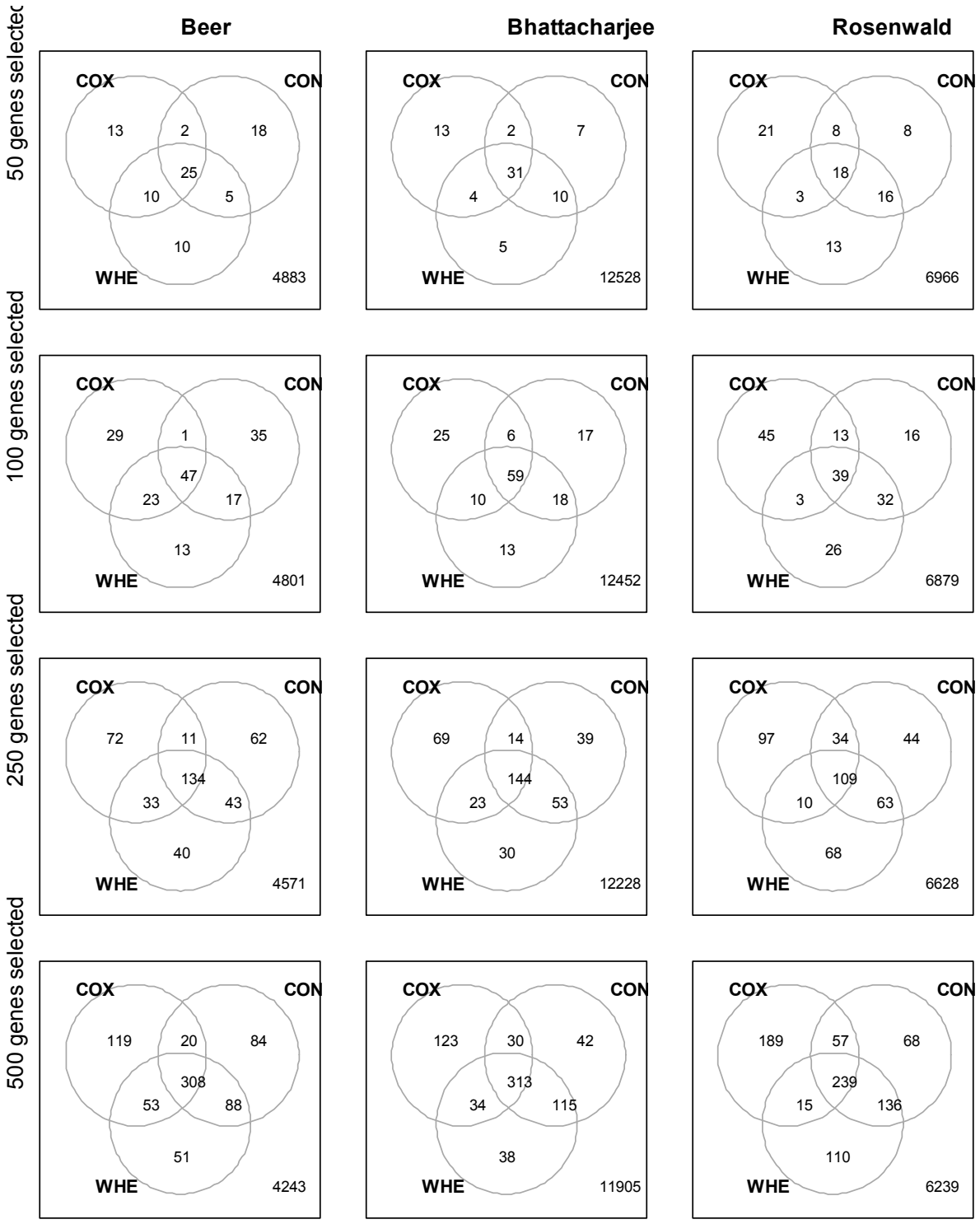
			Data		
	Statistic	**Method**	**Beer**	**Bhattacharjee**	**Rosenwald**
Select 50 genes with each method	$\hat{c}'_{+(50)}$	COX	0.681	0.551	0.563
	$\hat{c}'_{+(50)}$	WHE	0.681	0.556	0.574
	$\hat{c}'_{+(50)}$	CON	0.681	0.556	0.570
	FDR_{50}	*COX*	*0.297*	*1.216*	*0.339*
	FDR_{50}	*WHE*	*0.448*	*0.971*	*0.693*
	FDR_{50}	*CON*	*0.696*	*1.129*	*0.270*
	# genes selected by	COX & WHE	35 (70%)	35 (70%)	21 (42%)
	# genes selected by	COX & CON	27 (54%)	33 (66%)	26 (52%)
	# genes selected by	WHE & CON	30 (60%)	41 (82%)	34 (68%)
	# genes selected by	COX, WHE & CON	25 (50%)	31 (62%)	18 (36%)
Select 100 genes with each method	$\hat{c}'_{+(100)}$	COX	0.656	0.546	0.556
	$\hat{c}'_{+(100)}$	WHE	0.658	0.55	0.566
	$\hat{c}'_{+(100)}$	CON	0.661	0.551	0.563
	FDR_{100}	*COX*	*0.349*	*1.186*	*0.387*
	FDR_{100}	*WHE*	*0.421*	*0.947*	*0.737*
	FDR_{100}	*CON*	*0.721*	*1.039*	*0.336*
	# genes selected by	COX & WHE	70 (70%)	69 (69%)	42 (42%)
	# genes selected by	COX & CON	48 (48%)	65 (65%)	52 (52%)
	# genes selected by	WHE & CON	64 (64%)	77 (77%)	71 (71%)
	# genes selected by	COX, WHE & CON	47 (47%)	59 (59%)	39 (39%)
Select 250 genes with each method	$\hat{c}'_{+(250)}$	COX	0.628	0.539	0.548
	$\hat{c}'_{+(250)}$	WHE	0.626	0.543	0.557
	$\hat{c}'_{+(250)}$	CON	0.630	0.544	0.553
	FDR_{250}	*COX*	0.389	1.053	0.383
	FDR_{250}	*WHE*	0.492	0.841	0.721
	FDR_{250}	*CON*	0.845	0.923	0.369
	# genes selected by	COX & WHE	167 (67%)	167 (67%)	119 (48%)
	# genes selected by	COX & CON	145 (58%)	158 (63%)	143 (57%)
	# genes selected by	WHE & CON	177 (71%)	197 (79%)	172 (69%)
	# genes selected by	COX, WHE & CON	134 (54%)	144 (58%)	109 (44%)

			Data		
	Statistic	**Method**	**Beer**	**Bhattacharjee**	**Rosenwald**
Select 500 genes with each method	$\hat{c}'_{+(500)}$	COX	0.607	0.534	0.542
	$\hat{c}'_{+(500)}$	WHE	0.604	0.537	0.550
	$\hat{c}'_{+(500)}$	CON	0.607	0.538	0.545
	FDR_{500}	*COX*	*0.437*	*0.996*	*0.383*
	FDR_{500}	*WHE*	*0.556*	*0.834*	*0.702*
	FDR_{500}	*CON*	*0.870*	*0.892*	*0.384*
	# genes selected by	COX & WHE	361 (72%)	347 (69%)	254 (51%)
	# genes selected by	COX & CON	328 (66%)	343 (69%)	296 (59%)
	# genes selected by	WHE & CON	396(79%)	428 (86%)	375 (75%)
	# genes selected by	COX, WHE & CON	308 (62%)	313 (63%)	239 (48%)

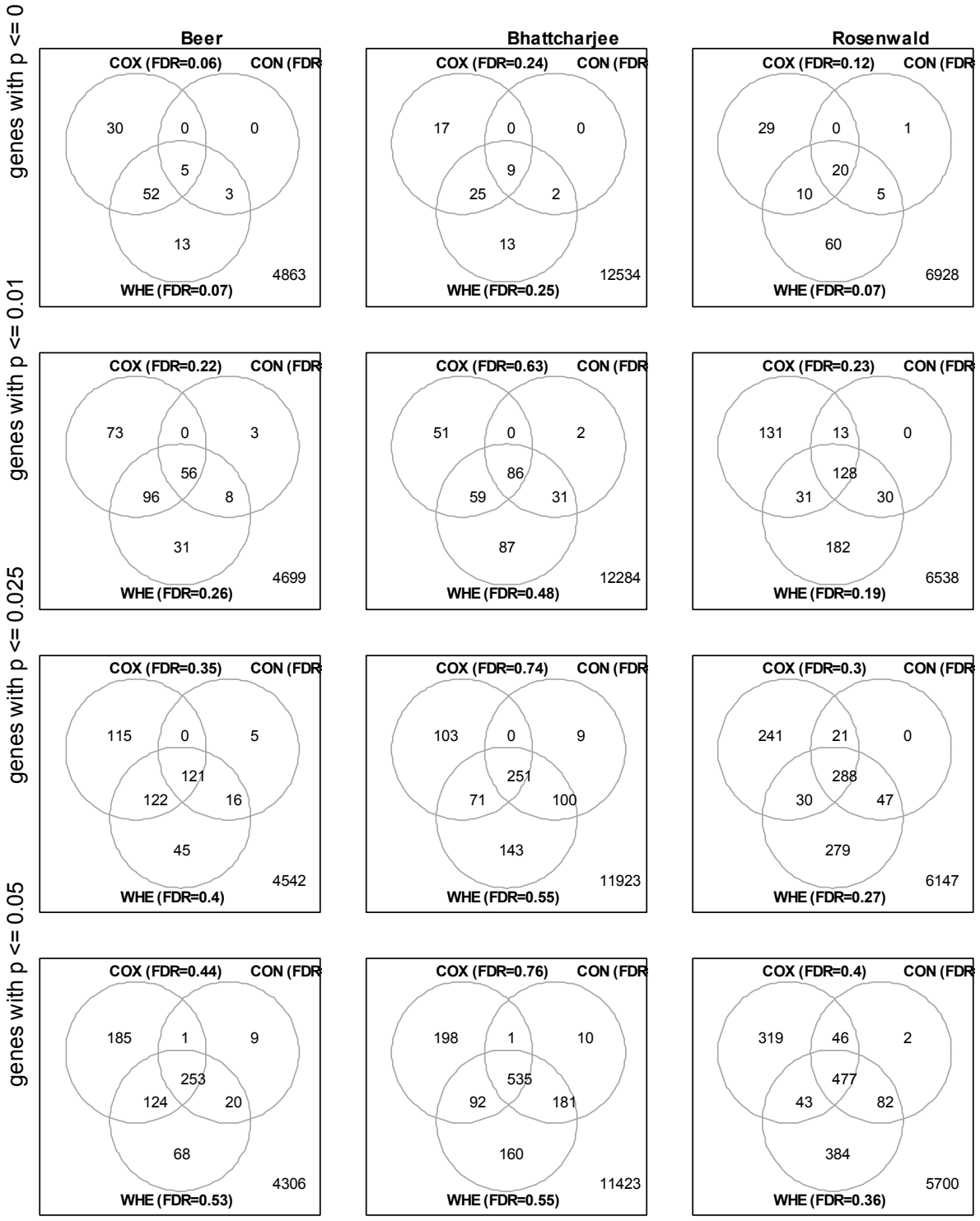
Web-Table 2: Results of analyses by Cox regression (COX), weighted Cox regression (WHE), concordance regression (CON) of three real-life data sets. In the general information section the table summarises the number of observations, events and genes; the number of genes with $\hat{c}' < 0.5$ in absolute numbers and percentages ('# of genes with $\hat{c}' < 0.5$'); the range of $\hat{c}'$ and the correlation of $\hat{c}'$ between two methods.
m top-ranked genes (sorted by $\hat{c}'_{+}$) are selected and the resulting gene lists are summarized with the $\hat{c}'_{+}$ value of the m^{th} gene ('$\hat{c}'_{+m}$'), the false discovery rate ('FDR_m') and the number of genes selected in common by COX, WHE and CON. *m* was set to 50, 100, 250 and 500.

Web-Figure 3 and 4 show Venn diagrams for the three analyzed data sets if genes are selected either by their absolute effect sizes $\hat{c}'_{+}$ or by their respective *p*-values. Irrespective of the data set and the number of selected genes WHE and CON show the highest agreement in gene selection; approximately 70% for the Beer and the Rosenwald data und approximately 80% for the Bhattacharjee data.

The agreement of genes selected by COX, WHE and CON is larger if *p*-values are used for gene selection (Web-Figure 4) instead of the absolute effect size $\hat{c}'_{+}$.



Web-Figure 3: Venn diagrams for three analyzed data sets, if the 50, 100, 250 or 500 genes with the largest absolute effect size $\hat{c}'_+$ are selected.

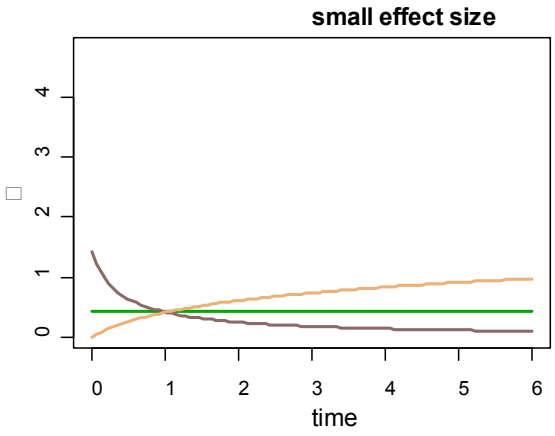


Web-Figure 4: Venn diagrams for three analyzed data sets, if genes with $p \leq 0.001, 0.01, 0.025$ or 0.05 are selected. The numbers in brackets is the step-up FDR (Benjamini and Hochberg, 1995).

4. Simulation study

We evaluated COX, WHE and CON by simulating trials assessing the association of gene expression with survival. The first series of simulations aimed at comparing the methods in univariate models considering expression of only one gene at the same time ('univariate evaluation'). These simulations should reveal differences of the methods in estimating the concordance probability c' as defined in the previous section under PH and various assumptions of NPH. A second series simulated typical gene expression studies, and we considered a large number of genes with partly correlated expression competing for selection in the same study ('multivariate evaluation').

In this series of simulations, we assumed that log gene expression values follow a standard normal distribution, and that survival time y follows a Weibull distribution with shape parameter $a = 2$ and scale parameter $b = 0.5$. The survival times were simulated by first generating uniformly distributed random numbers u from U[0,1], and inserting them into $y = -\log(u)^{1/a} / b$. Gene expression was generated from a standard normal distribution and linked to survival time by applying the algorithm of MacKenzie and Abrahamowicz (2002). This algorithm is described in detail in the appendix of this supplementary material. We made three assumptions on time-dependency, three assumptions on the strength of effects, and also three assumptions on presence and amount of censoring, which led to 27 investigated scenarios. In each scenario we simulated 2000 data sets of 200 observations each. For time dependency we considered PH with $\beta(t) = \beta_0$, CH with a time-dependent log hazard ratio of $\beta(t) = \beta_0 \left[1 + 2.88 / (1 + 5t)\right]$, and DH with $\beta(t) = \beta_0 (1 + 1.86\, t)$. β_0 was set to achieve pre-defined values for c' of 0.60 ('small' effect size), 0.66 ('medium' effect size) and 0.80 ('large' effect size). Under PH, these choices correspond to β_0 values of log(1.5), log(2) and log(4). Web-Figure 5 shows the resulting relationship of β and time for different scenarios of the univariate evaluation.


small effect size
0
1
2
3
4
0
1
2
3
4
5
6
time

To simulate censoring we drew a uniformly distributed follow-up time f from U[0, τ] and defined the observed survival time $t = \min(y, f)$ with status indicator $d = \mathrm{I}(f > y)$. We iteratively determined τ to obtain proportions of censored times of 33% and 67%.

When censoring is combined with time-dependent effects a part of the observed bias can be attributed to the discrepancy of the population value of c' given follow-up is restricted to a maximum time τ compared to the unrestricted c'. Since the relationship of β_0 and the 'follow-up-restricted population value of c'' cannot be analytically determined we had to employ simulation for the computation of the 'follow-up-restricted population value of c'', similarly to the description given above but omitting all pairs where $t_i > \tau$ and $t_j > \tau$. Across all scenarios, the largest observed discrepancy between the population value of c' and the 'follow-up-restricted population value of c'' was 0.023.

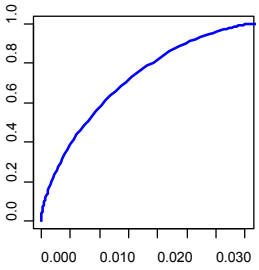
Web-Figure 6 shows boxplots of the estimates of c' by COX, WHE and CON.

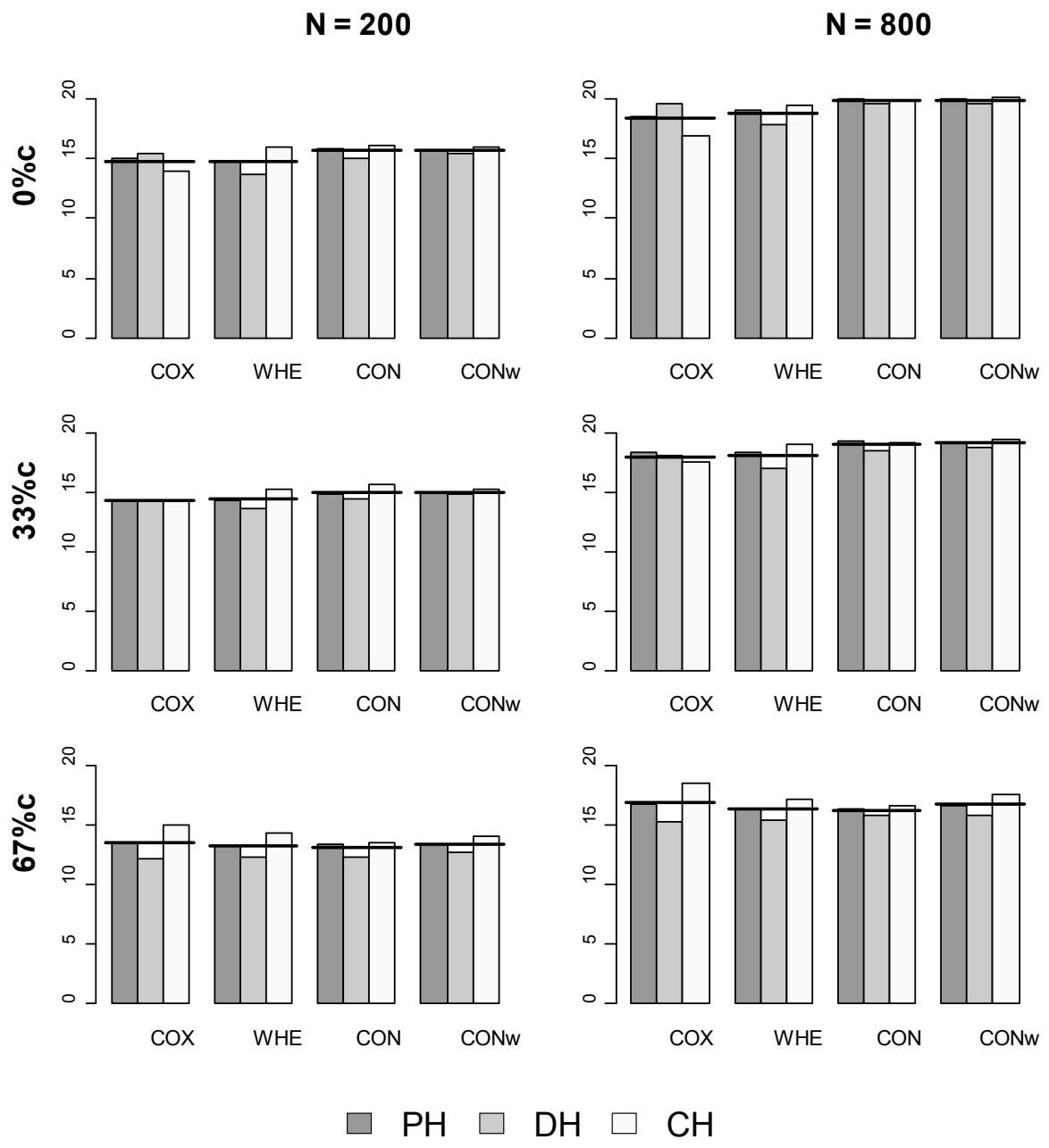
The aim of the second series of simulations was to see how the methods compare in selecting those genes which truly are related to survival, if a large number of genes are competing for selection. We simulated gene expression of $p = 5000$ features according to the scheme outlined by Binder and Schumacher (2008) and assumed that only the first 72 genes had an additive effect on the log hazard, with an equal number of 24 genes exhibiting PH, CH and DH. From each group, we chose 8 genes to have a 'large' effect size and 16 genes to have a 'small' effect size. As in the univariate simulation, we simulated survival times from a Weibull (2, 0.5) distribution with distribution function denoted by $F_W(t)$. We linked gene expression data to survival times, assuming that the hazard of individual *i* at time *t* is $\lambda_i(t) = \lambda_0(t)\exp\left[\sum_{g=1}^{p} x_{ig}\beta_g(t)\right]$. The time-dependent log hazard ratio of gene *g* was defined as $\beta_g(t) = \beta_0$ in case of PH, $\beta_g(t) = \beta_0\left[1 + 2.88/(1+5t)\right]$ for CH, and $\beta_g(t) = \beta_0(1+1.86\,t)$ for DH. The constants β_0 were set such that average regression effects $\overline{\beta} = \int \beta_g(t)dF_W(t)$ of 0.4 ('large' effect size) and 0.2 ('small' effect size) resulted. For each combination of censoring (0%, 33%, 67%) and sample size (200, 800) we generated 200 data sets and assessed the variability of results. The data generation algorithm is outlined in the appendix of this supplementary material.

Each data set was analyzed using COX, WHE and CON and for each gene c' was estimated. Genes were ranked by $\hat{c}_+^{'}$ and the *m* top genes were considered 'selected'. We tried various choices for *m*.

'Long survivors' in a data set may obtain very large weights in CON, resulting in extremely unequal contributions to the likelihood. To address this issue, the complete analysis was repeated with all CON weights truncated at the 95th percentile (CONw).

The true positive rates (TPR) resulting from varying the number of selected genes and the false positive rates (FPR) are contrasted in Web-Figure 7. TPR was defined as the rate of selected genes among the genes associated with survival, and FPR was the rate of false positive genes among those selected. Again, we notice advantages of CON under no and 33% censoring, and slightly less pronounced gains of COX under 67% censoring. The TPR is higher with a sample size of 800 compared to 200, but the overall tendencies do not change.


0.0
0.2
0.4
0.6
0.8
1.0
0.000
0.010
0.020
0.030



Web-Figure 8: Comparison of the average number of correctly selected genes by the type of time-dependency for data with 200 and 800 observations. The solid horizontal lines indicate the average over the three types of time-dependence. COX, Cox regression, WHE, weighted Cox regression; CON, concordance regression; CONw, concordance regression with truncation of weights at the 95th percentile; %c, percent censoring; N, sample size; PH, proportional hazards; DH, diverging hazards; CH, converging hazards.

The complete results of the multivariate evaluation are given below and are organized in the following Web-Figures and Web-Tables:

No.	Web-Figures and Web-Tables	Sample size
1	# of correctly selected genes if 100 genes are selected	200 + 800
2	# of correctly selected genes id 72† genes are selected	200 + 800
3	# of correctly selected genes if 24‡ genes are selected	200 + 800
4	FDR versus TPR for COX, WHE and CON	200 + 800
5	FDR versus TPR for CON and CONw	200
6	FDR versus TPR for CON and CONw	800
7	p-values from paired t-tests for method comparisons corresponding to Table 1	200 + 800

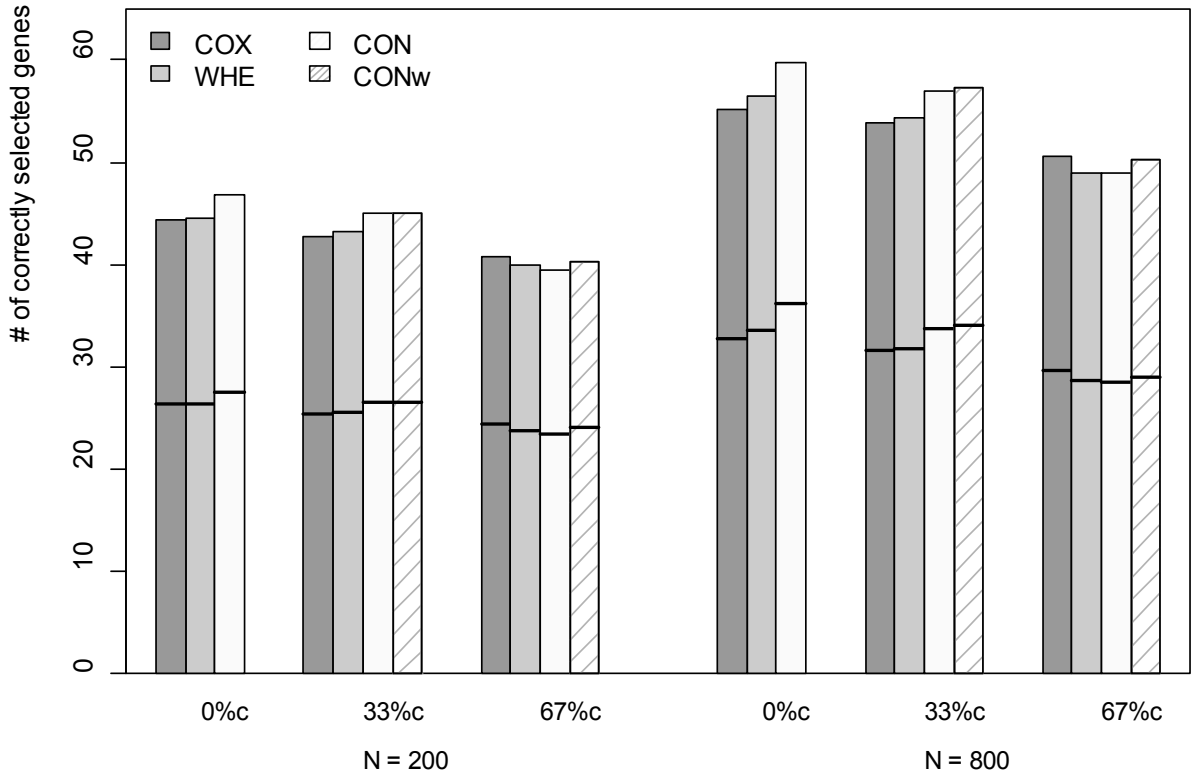
The average number of correctly selected genes under various censoring proportions and sample sizes is graphically compared in Nos. 1 to 3. The true positive rate is highest for CON in scenarios with no or 33% censoring, while COX outperforms WHE and CON at 67% censoring. For concordance regression, weight truncation (CONw) compensates the loss of efficiency with 33 or 67% censoring. This compensation even improves with higher sample size. Moving from a sample size of 200 to 800, the number of correctly selected genes increases by approximately 35% for CON (when 72 genes are selected) and by approximately 30% for COX and WHE (when 72 genes are selected).

Nos. 4 to 6 show the false positive rate (FPR) versus the true positive rate (TPR) for various censoring proportions and sample sizes and confirm the results of Nos. 1 to 3. Gains in efficiency if 800 instead of 200 observations are available are approximately the same in COX, WHE and CON. In case of censoring the application of CONw can additionally increase efficiency (Nos. 9 to 15).

† This is the number of genes with an assumed effect on survival.
‡ This is the number of genes with an assumed strong effect on survival.

Multivariate Evaluation No. 1

[Plot and table index]



Web-Figure 9: Average number of correctly selected genes by Cox (COX), weighted Cox (WHE) and concordance (CON) regression from the multivariate evaluation when 100 genes with the largest absolute effect size are selected. For 33 and 67% censoring results of concordance regression with truncation of weights at the 95th percentile (CONw) are additionally included. We assumed that 72 genes had an additive effect on the log hazard (48 with 'small' and 24 with 'large' effect size), with an equal number of 24 genes exhibiting proportional, diverging and converging hazards. Lower and upper parts of each bar correspond to correctly selected genes with 'small' and 'large' effect sizes, respectively. %c, percent censored; N, sample size.

N	hazard	0%c		33%c		67%c	
		small ES	large ES	small ES	large ES	small ES	large ES
200	PH	8.9/8.7/9.3	6.1/6.1/6.5	8.6/8.6/8.8	5.7/5.7/6.2	8.1/8.0/8.1	5.4/5.3/5.3
	DH	9.0/7.9/8.7	6.4/5.8/6.3	8.4/8.0/8.8	5.9/5.6/6.1	7.4/7.4/7.6	4.8/5.0/5.1
	CH	8.5/9.6/9.7	5.5/6.4/6.5	8.4/8.9/8.9	5.9/6.4/6.3	8.9/8.5/8.4	6.2/5.8/5.7
	subtotal	26.4/26.3/27.6	18.0/18.3/19.3	25.4/25.5/26.5	17.4/17.7/18.6	24.4/23.8/24.1	16.4/16.1/16.1
	*total**	44.4/44.6/**46.9**		42.8/43.2/**45.1**		**40.8**/39.9/40.2	
800	PH	11.0/11.4/12.2	7.6/7.7/7.9	10.8/10.8/11.4	7.5/7.5/7.8	9.8/9.6/9.7	7.0/6.8/7.0
	DH	12.0/10.6/11.9	7.7/7.3/7.7	10.6/9.8/11.0	7.5/7.2/7.7	9.0/9.0/9.2	6.3/6.4/6.7
	CH	9.7/11.6/12.0	7.2/7.9/7.9	10.2/11.3/11.6	7.3/7.7/7.8	10.9/10.0/10.1	7.6/7.2/7.5
	subtotal	32.7/33.6/36.2	22.5/22.9/23.5	31.6/31.8/34.0	22.3/22.5/23.3	29.6/28.6/29.0	21.0/20.3/21.2
	*total**	55.2/56.5/**59.7**		53.9/54.3/**57.3**		50.6/48.9/50.2	

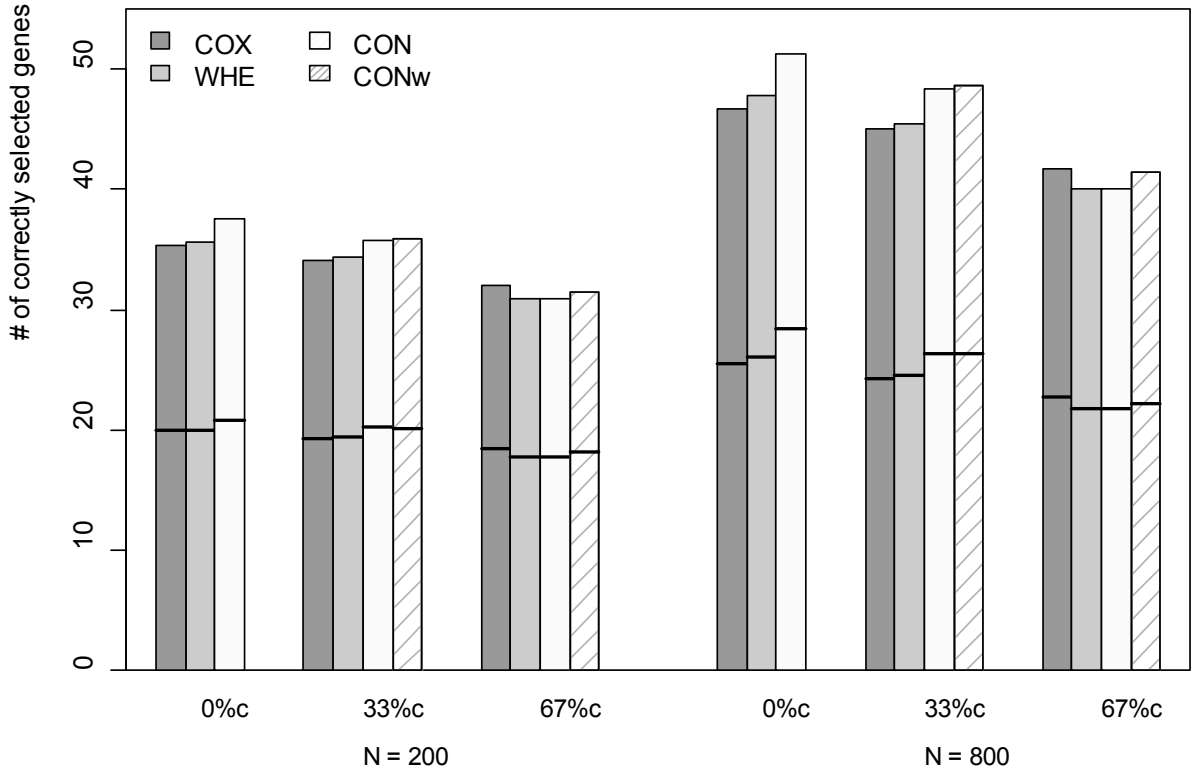
Web-Table 3: Average number of true positive genes in 200 simulated data sets selected by Cox/weighted Cox/concordance regression. In case of censoring concordance regression with truncation of weights at the 95th percentile is applied. The total number of selected genes was 100 for each method and each scenario. The effect sizes were set to 'large' for 24 and to 'small' for 48 out of 5000 candidate genes. PH, proportional hazards; CH, converging hazards; DH, diverging hazards; %c, percent censored; ES, effect size; N, sample size;
* The significantly ($p<0.01$) highest total number of true positive genes is set in boldface.

N	hazard	33%c		67%c	
		small ES	large ES	small ES	large ES
200	PH	8.8/8.8	6.1/6.2	8.0/8.1	5.4/5.3
	DH	8.5/8.8	6.0/6.1	7.3/7.6	5.0/5.1
	CH	9.2/8.9	6.4/6.3	8.1/8.4	5.5/5.7
	subtotal	26.5/26.5	18.5/18.6	23.5/24.1	15.9/16.2
	total	45/45.1		39.4/40.3	
800	PH	11.5/11.4	7.8/7.8	9.6/9.7	6.8/7.0
	DH	10.8/11.0	7.7/7.7	9.2/9.2	6.6/6.7
	CH	11.4/11.6	7.8/7.8	9.7/10.1	7.0/7.5
	subtotal	33.8/34.0	23.2/23.3	28.5/29.0	20.4/21.2
	total	57/57.3		48.9/51.2	

Web-Table 4: Average number of true positive genes in 200 simulated data sets selected by concordance regression/concordance regression with truncation of weights at the 95th percentile. The total number of selected genes was 100 for each method and each scenario. The effect sizes were set to 'large' for 24 and to 'small' for 48 out of 5000 candidate genes. PH, proportional hazards; CH, converging hazards; DH, diverging hazards; %c, percent censored; ES, effect size; N, sample size.

Multivariate Evaluation No. 2

[Plot and table index]



Web-Figure 10: Average number of correctly selected genes by Cox regression (COX), weighted Cox regression (WHE) and concordance regression (CON) from the multivariate evaluation when 72 genes with the largest absolute effect size are selected. For 33 and 67% censoring results of concordance regression with truncation of weights at the 95th percentile (CONw) are additionally included. We assumed that 72 genes had an additive effect on the log hazard (48 with 'small' and 24 with 'large' effect size), with an equal number of 24 genes exhibiting proportional, diverging and converging hazards. Lower and upper parts of each bar correspond to correctly selected genes with 'small' and 'large' effect sizes, respectively. %c, percent censored; N, sample size.

N	Hazard	0%c		33%c		67%c	
		small ES	large ES	small ES	large ES	small ES	large ES
200	PH	6.8/6.7/7.0	5.2/5.1/5.6	6.5/6.5/6.6	4.8/4.8/5.3	6.2/5.9/6.1	4.5/4.4/4.4
	DH	6.8/5.8/6.6	5.6/4.9/5.4	6.3/6.1/6.7	5.0/4.7/5.1	5.3/5.3/5.5	3.8/3.8/4.1
	CH	6.3/7.5/7.2	4.6/5.5/5.7	6.5/6.8/6.7	5.0/5.5/5.3	6.9/6.6/6.6	5.3/4.9/4.8
	subtotal	19.9/20.0/20.8	15.4/15.6/16.8	19.3/19.4/20.1	14.8/14.9/15.8	18.4/17.8/18.2	13.6/13.1/13.3
	*total**	35.3/35.6/**37.6**		34.1/34.3/**35.9**		32.0/30.9/31.5	
800	PH	8.6/8.8/9.6	7.2/7.2/7.7	8.3/8.4/8.8	6.9/7.0/7.5	7.5/7.2/7.6	6.3/6.1/6.4
	DH	9.5/8.0/9.4	7.4/6.8/7.4	8.1/7.2/8.5	7.0/6.7/7.3	6.5/6.8/6.7	5.3/5.4/5.9
	CH	7.5/9.2/9.5	6.6/7.6/7.8	7.8/8.8/9.1	6.8/7.3/7.5	8.9/7.9/7.9	7.2/6.7/6.9
	subtotal	25.5/26.1/28.4	21.2/21.7/22.9	24.3/24.5/26.3	20.7/21.0/22.3	22.8/21.8/22.2	18.9/18.2/19.2
	*total**	46.7/47.8/**51.3**		45.0/45.5/**48.6**		41.7/40.0/41.4	

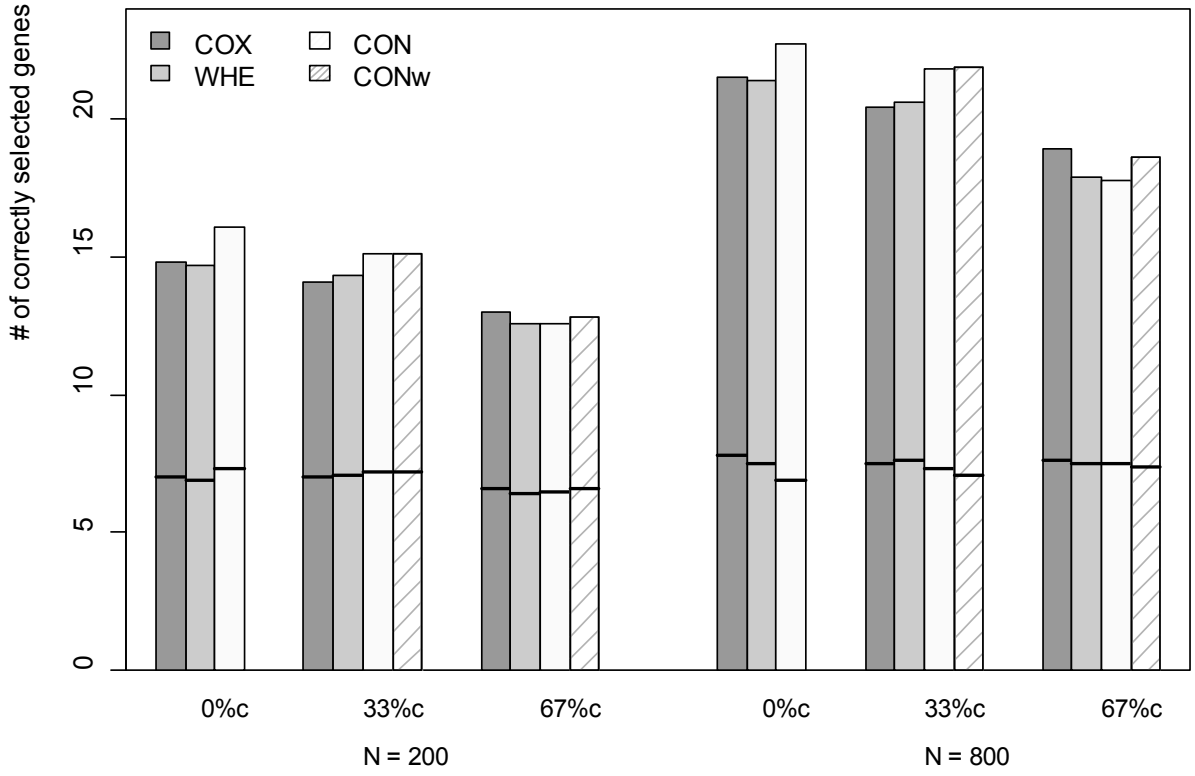
Web-Table 5: Average number of true positive genes in 200 simulated data sets selected by Cox/weighted Cox/concordance regression. In case of censoring concordance regression with truncation of weights is applied at the 95th percentile. The total number of selected genes was 72 for each method and each scenario. The effect sizes were set to 'large' for 24 and to 'small' for 48 out of 5000 candidate genes. PH, proportional hazards; CH, converging hazards; DH, diverging hazards; %c, percent censored; ES, effect size; N, sample size;
***** The significantly ($p<0.01$) highest total number of true positive genes is set in boldface.

N	hazard	33%c		67%c	
		small ES	large ES	small ES	large ES
200	PH	6.6/6.6	5.1/5.3	6.2/6.1	4.5/4.4
	DH	6.5/6.7	5.0/5.1	5.4/5.5	4.0/4.1
	CH	7.0/6.7	5.5/5.3	6.2/6.6	4.7/4.8
	subtotal	20.2/20.1	15.6/15.8	17.8/18.2	13.1/13.3
	total	35.8/35.9		30.9/31.5	
800	PH	9.0/8.8	7.3/7.5	7.3/7.6	6.1/6.4
	DH	8.5/8.5	7.3/7.3	6.9/6.7	5.6/5.9
	CH	8.8/9.1	7.4/7.5	7.6/7.9	6.5/6.9
	subtotal	26.3/26.3	22.1/22.3	21.8/22.2	18.2/19.2
	total	48.3/48.6		40.0/41.4	

Web-Table 6: Average number of true positive genes in 200 simulated data sets selected by concordance regression/concordance regression with truncation of weights at the 95th percentile. The total number of selected genes was 72 for each method and each scenario. The effect sizes were set to 'large' for 24 and to 'small' for 48 out of 5000 candidate genes. PH, proportional hazards; CH, converging hazards; DH, diverging hazards; %c, percent censored; ES, effect size; N, sample size.

Multivariate Evaluation No. 3

[Plot and table index]



Web-Figure 11: Average number of correctly selected genes by Cox (COX), weighted Cox (WHE) and concordance (CON) regression from the multivariate evaluation when 24 genes with the largest absolute effect size are selected. For 33 and 67% censoring results of concordance regression with truncation of weights at the 95th percentile (CONw) are additionally included. We assumed that 72 genes had an additive effect on the log hazard (48 with 'small' and 8 with 'large' effect size), with an equal number of 24 genes exhibiting proportional, diverging and converging hazards. Lower and upper parts of each bar correspond to correctly selected genes with 'small' and 'large' effect sizes, respectively. %c, percent censored; N, sample size.

N	hazard	0%c small ES	0%c large ES	33%c small ES	33%c large ES	67%c small ES	67%c large ES
200	PH	2.6/2.5/2.6	2.6/2.4/2.9	2.3/2.2/2.4	2.2/2.3/2.6	2.2/2.2/2.2	2.1/2.1/2.0
	DH	2.4/1.8/2.2	3.0/2.3/2.8	2.2/2.1/2.4	2.4/2.0/2.5	1.8/1.8/1.9	1.6/1.7/1.9
	CH	2.0/2.6/2.5	2.2/3.1/3.1	2.4/2.7/2.4	2.5/2.9/2.8	2.6/2.4/2.5	2.8/2.5/2.4
	subtotal	7.0/6.9/7.3	7.8/7.8/8.8	7.0/7.1/7.2	7.1/7.2/7.9	6.6/6.4/6.5	6.4/6.2/6.1
	*total**	14.8/14.7/**16.1**		14.1/14.3/**15.1**		**13.0**/12.6/12.6	
800	PH	2.8/2.6/2.4	4.6/4.5/5.3	2.5/2.5/2.5	4.2/4.2/4.8	2.4/2.4/2.5	3.6/3.4/3.7
	DH	2.9/1.9/2.0	5.3/3.7/4.8	2.5/2.1/2.2	4.4/3.6/4.7	1.8/2.1/2.1	2.5/2.6/2.9
	CH	2.1/3.0/2.5	3.8/5.7/5.6	2.4/3.0/2.4	4.2/5.2/5.3	3.3/3.0/2.9	5.2/4.4/4.6
	subtotal	7.8/7.5/6.9	13.7/13.9/15.8	7.5/7.6/7.1	12.9/13.0/14.8	7.6/7.5/7.5	11.3/10.4/11.2
	*total**	21.5/21.4/**22.7**		20.4/20.6/**21.9**		18.9/17.9/18.7	

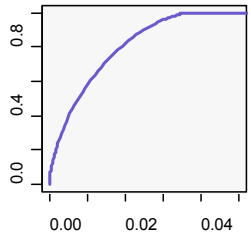
Web-Table 7: Average number of true positive genes in 200 simulated data sets selected by Cox/weighted Cox/concordance regression. In case of censoring concordance regression with truncation of weights is applied at the 95th percentile. The total number of selected genes was 24 for each method and each scenario. The effect sizes were set to 'large' for 24 and to 'small' for 48 out of 5000 candidate genes. PH, proportional hazards; CH, converging hazards; DH, diverging hazards; %c, percent censored; ES, effect size; N, sample size;
* The significantly ($p<0.01$) highest total number of true positive genes is set in boldface.

N	hazard	33%c small ES	33%c large ES	67%c small ES	67%c large ES
200	PH	2.4/2.4	2.5/2.6	2.1/2.2	2.0/2.0
	DH	2.2/2.4	2.5/2.5	2.0/1.9	1.8/1.9
	CH	2.6/2.4	2.9/2.8	2.4/2.5	2.2/2.4
	subtotal	7.2/7.2	7.9/7.9	6.5/6.6	6.1/6.2
	total	15.1/15.1		12.6/12.8	
800	PH	2.5/2.5	4.7/4.8	2.5/2.5	3.4/3.7
	DH	2.2/2.2	4.6/4.7	2.3/2.1	2.8/2.9
	CH	2.6/2.4	5.2/5.3	2.6/2.9	4.0/4.6
	subtotal	7.3/7.1	14.5/14.8	7.5/7.4	10.3/11.2
	total	21.8/21.8		17.8/18.6	

Web-Table 8: Average number of true positive genes in 200 simulated data sets selected by concordance regression/concordance regression with truncation of weights at the 95th percentile. The total number of selected genes was 24 for each method and each scenario. The effect sizes were set to 'large' for 24 and to 'small' for 48 out of 5000 candidate genes. PH, proportional hazards; CH, converging hazards; DH, diverging hazards; %c, percent censored; ES, effect size; N, sample size.

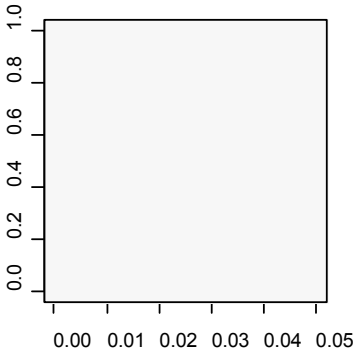
Multivariate Evaluation No. 4

[Plot and table index]



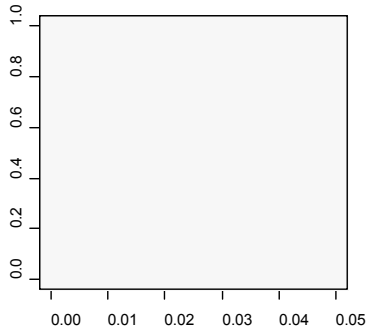
Multivariate Evaluation No. 5

[Plot and table index]



Multivariate Evaluation No. 6

[Plot and table index]



p-values from paired t-tests for method comparisons corresponding to Table 1

[Plot and table index]

N	Methods	0%c	33%c	67%c
200	COX-WHE	0.166	0.146	1.8E-13
	COX-CON	2.0E-29	1.3E-20	0.039
	WHE-CON	2.8E-26	3.25E-7	5.8E-7
800	COX-WHE	1.2E-9	5.6E-4	1.6E-29
	COX-CON	1.9E-156	1.5E-101	0.551
	WHE-CON	3.2E-90	1.9E-72	3.3E-4

Web-Table 9: p-values for pairwise comparisons between Cox (COX), weighted Cox (WHE) and concordance (CON) regression corresponding to Table 1. In case of censoring concordance regression with truncation of weights at the 95th percentile is applied. COX, Cox regression; WHE, weighted Cox regression; CON, concordance regression; %c, percent censored; N, sample size.

5. Abbreviations

c	Effect size measure; $c = P(T_1 < T_0)$, with T_1 and T_0 as randomly chosen survival times of group 1 and 0.
c'	Generalization of c to continuous data
c'_+	Absolute effect size; $c'_+ = 0.5 + \lvert c' - 0.5 \rvert$,
$\hat{\pi}_0$	Proportion of genes not linked to survival
CH	Converging hazard
CON	Concordance regression
CONw	Concordance regression with truncation of weights
COX	Cox regression
DH	Diverging hazard
ES	Effect size
FDR	False discovery rate
FP	False positives
FPR	False positive rate
HR	Hazard ratio
N	Sample size
NPH	Non-proportional hazard
PH	Proportional hazard
R	Correlation
TPR	True positive rate
WHE	Weighted Cox regression
%c	Percent censoring

6. References

Beer,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816-824.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B.*, **57**, 289-300.

Bhattacharjee,A. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA*, **98**, 13790-13795.

Binder,H. and Schumacher,M. (2008) Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. *Stat. Appl. Genet. Mol. Biol.*, **7**, Article 12.

MacKenzie,T. and Abrahamowicz,M. (2002) Marginal and hazard ratio specific random data generation: Applications to semi-parametric bootstrapping. *Stat. Comput.*, **12**, 245-252.

Rosenwald,A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, **346**, 1937-1947.

Tusher,V.G., Tibshirani,R.J. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116-5121.

Appendix: Data generating algorithm

Data generation for the simulation study followed the algorithms outlined by MacKenzie and Abrahamowicz (2002) and Binder and Schumacher (2008).

To simulate one sample of size *N*, this algorithm takes the following steps:

1. Generate *N* survival times $y_i,\ i=1,\ldots,N$, from the Weibull(2, 0.5)-distribution as follows:
   a. Draw a random number v_i from the uniform U(0,1)-distribution.
   b. Set $y_i = \frac{(-\log v_i)^{1/2}}{0.5}$.
2. Generate *N* follow-up times $z_i,\ i=1,\ldots,N$, from a uniform distribution between 0 and τ. τ was determined such that from step 3 censoring proportions of 0 $(\tau=\infty)$, 33% or 67% resulted.
3. Sort the *N* tuples (t_i, d_i) such that $t_i < t_{i+1}$, where $t_i = \min(y_i, z_i)$ and $d_i = I(y_i \le z_i)$.
4. For each subject *j* ($j=1,\ldots,N$), draw gene expression values x_{jg} ($g=1,\ldots,p$) from a standard normal distribution.
   a. In case of the univariate simulation, *p*=1.
   b. In case of the multivariate simulation, we set *p*=5000. For each gene *g* expression values are determined by

$$a_{jg} = \begin{cases} -1+\varepsilon_{jg} & \text{if } j \le 0.5n, g \le 0.05p \\ 1+\varepsilon_{jg} & \text{if j} > 0.5n, g \le 0.05p \\ 1.5\cdot I\left(u_{j1}<0.4\right)+\varepsilon_{jg} & \text{if } 0.05 < g \le 0.1p \\ 0.5\cdot I\left(u_{j2}<0.7\right)+\varepsilon_{jg} & \text{if } 0.1 < g \le 0.2p \\ 1.5\cdot I\left(u_{j3}<0.3\right)+\varepsilon_{jg} & \text{if } 0.2 < g \le 0.3p \\ \varepsilon_{jg} & \text{if } g > 0.3p \end{cases}$$

Resulting in a correlation of about
0.50
0.50
0.35
0.05
0.32
0

with $\varepsilon_{jg} \sim N(0,1)$ denoting a standard normally distributed error term, u_{jg} a uniform random variable in the range [0,1] and $I()$ denoting the indicator function, assuming the values 1 (if the argument is true) or 0 (if the argument is not true).

5. In this step we assign to each tuple (t_i, d_i), $i = 1, \ldots, N$, a gene expression vector $x_{i.}$ which is sampled from the N gene expression vectors $a_{j.}$, $j = 1, \ldots, N$, generated in the previous step. We start at the smallest observed survival time t_1 and the corresponding risk set $R_1 = \{1, \ldots, N\}$. For each $i = 1, \ldots, N$, sample a subject $j^* \in R_i$, whose gene expression vector $a_{j^*.}$ will be assigned to $x_{i.}$, as follows:

   a. If $d_i = 0$, then randomly sample j^* from the risk set R_i. Remove that subject from the risk set such that $R_{i+1} = R_i \quad \{j^*\}$.

   b. If $d_i = 1$, then sample j^* from the risk set R_i by assigning sampling probabilities proportional to $\exp\left[\sum_{g=1}^{p} x_{jg} \beta_g(t)\right]$ to the subjects $j \in R_i$. Set $R_{i+1} = R_i \quad \{j^*\}$.