BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 6 2010, pages 784–790
doi:10.1093/bioinformatics/btq035

Gene expression

Advance Access publication January 29, 2010

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

Daniela Dunkler, Michael Schemper and Georg Heinze*

Section of Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, 1090 Vienna, Austria

Associate Editor: Trey Ideker

**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 over-estimation 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.

**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:** georg.heinze@meduniwien.ac.at

**Supplementary Information:** Supplementary Data are available at *Bioinformatics* online.

Received on October 21, 2009; revised on January 14, 2010; accepted on January 22, 2010

## 1 INTRODUCTION

In recent years, many studies aimed at finding prediction models which link high-dimensional gene expression data to a survival outcome (e.g. Beer *et al.*, 2002; Bhattacharjee *et al.*, 2001; Rosenwald *et al.*, 2002). A comparative analysis of methods used in this context can be found in a paper by Bøvelstad *et al.* (2007). All these methods are extensions of the basic Cox proportional hazards regression model (COX) (Cox, 1972). This semi-parametric model assumes that the hazard rate $\lambda_i(t|x_i)$ of subject $i$ at time $t$ given a row vector $x_i$ of $\log_2$ gene expression measurements is of the form

$$\lambda_i(t|x_i)=\lambda_0(t)\exp(x_i\beta),$$

where $\lambda_0(t)$ is an unspecified baseline hazard function. Thus the elements of $\beta$ are log hazard ratios (HR) associated with a unit increase in $\log_2$ gene expression or a doubling of gene expression. COX assumes proportional hazards (PH), which means that we consider a constant effect of gene expression on survival over the whole period of follow-up. In a typical microarray study comprising a large number of genes the assumption of PH cannot be verified for each gene, though it is unlikely that PH hold for each gene. Consequently, ignoring the PH assumption and applying COX based methods could lead to under- or over-estimation for a considerable number of genes, meaning that some genes may be falsely declared important for predicting survival, while at the same time relevant genes are missed. If some genes exhibit non-proportional hazards (NPH) then their HRs, estimated while ignoring time-dependency, are not comparable to those of genes with PH or of genes exhibiting different patterns of NPH. Nonetheless, very few theoretical studies have considered the possibility of NPH and its consequences on gene selection or prediction (Xu *et al.*, 2005) and to our knowledge its occurrence has never been addressed in applied studies. By contrast, in classical applications a large body of literature exists which deals with COX under NPH (Abrahamowicz and MacKenzie, 2007; Collett, 2003; Marubini and Valsecchi, 1995) and the importance to cope with the possibility of NPH was often emphasized (Valsecchi *et al.*, 1996). With gene expression as predictor, this has been a neglected area of research.

Figure 1 and Supplementary Figure 1 show examples of genes with PHs, converging hazards (CH) and diverging hazards (DH) from the study of Bhattacharjee *et al.* (2001). For the gene with PH (left column) the correlation of scaled Schoenfeld residuals (Grambsch and Therneau, 1994) with the rank of time is close to 0. For the other genes the correlation is considerably larger indicating violation of the PH assumption. The middle column of Figure 1 shows a gene with CH, where the effect fades away with time, whereas the gene depicted in the right column exhibits DH, i.e. an effect increasing with time. In the Bhattacharjee study, genes with DH are found more often than genes with CH. NPH may arise from time-dependent effects of genes on survival, but could also result from model misspecification, e.g. from omitting a strong clinical covariate or another gene. This may particularly happen in univariate analyses.

### 1.1 Alternatives to COX in case of NPH

To cope with an apparent time-dependent effect of gene expression on survival, one may model the functional form of the time-dependency by an interaction of gene expression with arbitrary functions of time. Suitable functions of time could be found, e.g. by fractional polynomials or penalized regression (cf. Lehr and Schemper, 2007) or restricted cubic splines (Hess, 1994), or by a simple interaction with a monotonic function, e.g. the logarithm, of survival time (Ng'andu, 1997). In this article, we do not follow any of these approaches, because finding the best functional form of time-dependency for a huge number of predictors may lead to numerous problems like multicollinearity or multiple testing issues,

*To whom correspondence should be addressed.

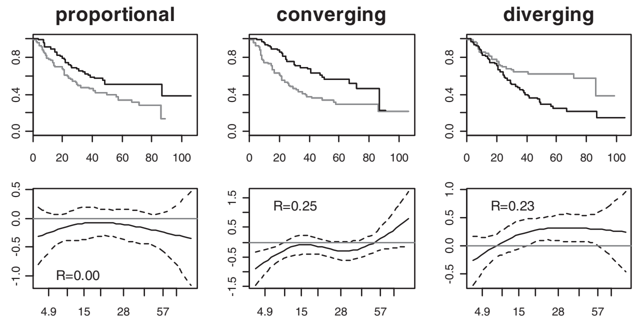

**Fig. 1.** Exemplification of proportional, converging and diverging hazards with three genes from the Bhattacharjee study. First row: Kaplan Meier estimates with gene expression dichotomized at the median versus time; second row: smoothed scaled Schoenfeld residuals and 95% confidence intervals versus rank of time. $R$, correlation of these residuals with rank of time.

to name just two. Other options to cope with NPH, like stratification or separate modeling for different time periods (Moreau *et al.*, 1985), are also not feasible with microarray data. Furthermore, all these options do not allow a comparison or ranking of genes with respect to their ability to predict short or long survival.

In this article, we propose a semi-parametric generalization of the well-known concordance probability as a summary measure of effect size suitable to rank genes when some of the genes may exhibit a time-dependent effect on survival. The concordance probability $c$ will be reviewed and its generalization $c'$ presented in Methods section, where we also introduce two methods to estimate $c'$: concordance regression (CON) and weighted Cox regression (WHE). In Results section we present results from a simulation study comparing the performance of COX, WHE and CON in estimating $c'$ under various conditions, and in producing gene lists in microarray experiments. We will also present results of analyses of three data sets, and close with a discussion in Section 4.

## 2 METHODS

### 2.1 The effect size measure c′ for continuous data

An intuitive non-parametric measure of separation of the survival distributions of two groups is the concordance probability $c=P(T_1<T_0)$, defined as the probability that a randomly chosen survival time $T_1$ from group $G_1$ is smaller than a randomly chosen survival time $T_0$ from group $G_0$. With uncensored data $c$ is equivalent to the non-parametric two-sample test statistics of Wilcoxon (1945) and Mann and Whitney (1947), and to the area under an ROC curve (Hanley and McNeil, 1982). Assuming PH, one can directly use the Cox regression coefficient $\hat{\beta}_C$ associated with a single binary covariate as an estimate of the log odds of $c$, i.e. $\hat{\beta}_C$ has the interpretation of $\log(c/(1-c))$. This interesting relationship is shown in detail in the Appendix. Under NPH this connection to $\hat{\beta}_C$ no longer applies, but $c$ still is a measure with an intuitive interpretation. However, as recently demonstrated, $c$ can be approximated also under NPH by using weighted estimation in Cox regression (Schemper *et al.*, 2009).

Since gene expressions are continuous rather than binary, the definition of $c$ has to be generalized to continuous data. Assume that $X$ denotes the $\log_2$ expression of some gene of interest. Then, such a generalization of $c$ could be defined as

$$c'=P\left(T_i<T_j|x_i=x_j+1\right),$$

where $T_i$ and $T_j$ are the survival times of randomly chosen subjects with $\log_2$ expressions $x_i$ and $x_j$, respectively. Since a one-unit increase in $X$ corresponds to a doubling of gene expression, $c'$ corresponds to the probability that survival time decreases if gene expression is doubled. Similarly, $\gamma=\log\left[c'/(1-c')\right]$ are the log odds that the survival time decreases if gene expression is doubled. Often, gene expression data are standardized to a common measure of spread (e.g. the standard deviation) across all genes and then our definition of $c'$ applies to a change of 1 SD. For convenience, we now assume that the log odds of concordance $\Gamma(x_i,x_j)=\text{logit}\left[P\left(T_i<T_j|x_i>x_j\right)\right]$ between two subjects with arbitrary $\log_2$ gene expression values $x_i$ and $x_j$ are proportional to $(x_i-x_j)$. This assumption corresponds to the linearity assumption of a PHs model, and implies that $\Gamma(x_i,x_j)/(x_i-x_j)=\gamma$ irrespective of the actual values of $x_i$ and $x_j$. Even under mild departures from this assumption, $\gamma$ may still be a useful summary measure if redefined as the expectation of $\Gamma(x_i,x_j)/(x_i-x_j)$ over all pairs of values $(x_i,x_j)$:

$$\begin{aligned}\gamma&=E_{(x_i,x_j)}\left[\Gamma\left(x_i,x_j\right)/\left(x_i-x_j\right)\right]\\&=\iint\Gamma\left(x_i,x_j\right)/\left(x_i-x_j\right)dF\left(x_i\right)dF\left(x_j\right)\end{aligned}$$

This quantity can be transformed into the generalized concordance probability by $c'=\exp(\gamma)/[1+\exp(\gamma)]$.

Our definition of $c'$ has some similarities with the concordance probability defined for time-to-event settings as $\text{CP}(X,T)=P\left(T_i<T_j|x_i>x_j\right)$ (Gönen, 2007, p. 89). $\text{CP}(X,T)$ is purely non-parametric and could also be used with gene expression data. Under PHs, Gönen and Heller (2005) have developed a modification of $\text{CP}(X,T)$ which is not sensitive to censoring. However, we prefer $c'$ here, because unlike $\text{CP}(X,T)$, $c'$ assumes a higher concordance probability with higher difference in gene expression.

### 2.2 Estimation of c′

*2.2.1 Concordance regression* We now propose to model $c'$ via its log odds $\gamma$ by $P\left(T_i<T_j|x_i>x_j\right)=\exp(x_i\gamma)/\left[(\exp(x_i\gamma)+\exp(x_j\gamma)\right]$. The associated log likelihood and its derivative can be written as

$$\ell(\gamma)=\sum_{(i,j)}\{x_i\gamma-\log[\exp(x_i\gamma)+\exp(x_j\gamma)]\},$$

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

where summation is over all available pairs $(i,j)$ such that $t_i<t_j$. These pairs $(i,j)$ will in the sequel be denoted as 'risk pairs' (as opposed to 'risk sets' in COX), and subjects may appear in multiple risk pairs. In our model the dependent variable is the concordance of the risk pair $(i,j)$ and hence, setting the first derivative of the log likelihood to zero yields a direct estimate of the log odds of concordance of $t_i<t_j$ related to a one-unit increase in $X$. Therefore, we denote this method as concordance regression. Our approach is semi-parametric as it does not require approximating or knowing the survivor function $S(T|X)$. Once an estimate $\hat{\gamma}$ has been computed, $\hat{c}'$ can be derived by

$$\hat{c}'=\exp(\hat{\gamma})/\{1+\exp(\hat{\gamma})\}.$$

Despite the continuous nature of gene expression, in practical problems ties in gene expression may occur. In this case, we omit risk pairs with $x_i=x_j$ from the likelihood, since they do not contribute information about $\gamma$ or $c'$.

In case of censoring, we omit all risk pairs where $t_i$ is censored, because it is not clear whether the true underlying survival time is less than $t_j$. Therefore, censoring leads to an overrepresentation of some subjects compared to others. In order to obtain an unbiased estimate of $c'$ despite this overrepresentation, we weight the risk pairs by their inverse sampling probabilities. Suitable weights are defined by

$$w(t_i)=[N(0)S(t_i)-1]/[N(t_i)-1]G(t_i)^{-1},$$

with $S(t)$ denoting the left continuous version of the Kaplan–Meier estimate of the survivor function at time $t$, $N(t)$ the number of patients at risk at $t$, and G($t$) denoting the probability to be still under follow-up at $t$, estimated by Kaplan–Meier but with the meaning of the status indicator reversed. The first term of the weight, $[N(0)S(t_i)-1]/[N(t_i)-1]$, restores the number of comparisons of a subject failing at time $t_i$ with subjects surviving that time that would have arisen had censoring not occurred. The second term of the weight, $G(t_i)^{-1}$, puts more weight on later event times compared to earlier times, and thus corrects the attenuation in observed events due to earlier censorship. It permits reconstructing the density of event times in the time range covered, i.e. till the last event. The weights $w(t_i)$ are introduced into the score function

$$\partial\ell(\gamma)/\partial\gamma=\sum_{(i,j)}D_{ij}w(t_i)\left[x_i-\frac{x_i\exp(x_i\gamma)+x_j\exp(x_j\gamma)}{\exp(x_i\gamma)+\exp(x_j\gamma)}\right],$$

where $D_{ij}$ is defined as 1 if $t_i<t_j$ and $t_i$ is uncensored, and 0 else. In a censored sample, the influence of the subjects on the likelihood is not equal; subjects with a long follow-up will contribute more information than those who are censored early. This unequal weighting will not bias point estimates unless, at the same time, censoring depends on gene expression and the effect of $\log_2$ gene expression is not linear. While the first assumption can be ruled out in most applications, linearity is a standard assumption also in COX, and thus does not distinguish our method from others. Since only the combined violation of both assumptions may lead to biased point estimates, our proposed estimate can be seen as a doubly robust estimate. However, unlike in COX, one cannot use the negative inverse of the second derivative of the likelihood as variance estimate, since summation is done over risk pairs and not over risk sets. Proper variance estimates can be obtained either by the jackknife or by a robust sandwich estimate (Lin and Wei, 1989; Therneau and Grambsch, 2000), but variance estimation will not be pursued here.

*2.2.2 Weighted Cox regression* Recently, Schemper *et al.* (2009) have shown that by introducing weights into the score function of Cox's partial log likelihood, an approximative estimate $\hat{\beta}_W$ of the log odds of concordance $\gamma$ is obtained that works well over a wide range of underlying values. The validity of the approximation is independent of the type of non-proportionality. The weights that are introduced are defined by $w(t_i)=S(t_i)G(t_i)^{-1}$, with $S(t_i)$ and $G(t_i)$ as defined above.

Another related approach was proposed by Xu and O'Quigley (2000). These authors also introduce weights into the score function, but their weights are defined by $w(t_i)=S(t_i)/N(t_i)$, which can be rewritten as $w(t_i)=G(t_i)^{-1}$. Their aim was to provide an average regression effect independent of the pattern of censoring. One advantage of the approach of Schemper *et al.* (2009) is the intuitive interpretation of $\hat{\beta}_W$ as log odds of concordance $\gamma$.

*2.2.3 Software* WHE has been implemented in an R package `coxphw` and a SAS macro `WCM` available at CRAN.r-project.org and http://www.muw.ac.at/msi/biometrie/programs, respectively. CON for estimation of $c'$ has also been implemented in an R package `concreg` and is available upon request from the authors.

### 2.3 Gene selection based on c′

Under PH, all three methods (COX, CON and WHE) will approximately supply similar estimates. Under NPH, however, we may expect differences between COX and CON or WHE, but similarity of the latter. We now assume that gene selection is done based on univariate regressions, and assume that from all candidate genes, the top-ranked are selected for further analysis. In our context, we rank genes by their absolute effect size $c'_+=0.5+|c'-0.5|$, estimated via $|\hat{\beta}_C|$, $|\hat{\beta}_W|$ or $|\hat{\gamma}|$ supplied by COX, WHE or CON, respectively. A threshold on $c'_+$ can be defined to produce a list of 'significant' genes, and the false discovery rate (FDR) associated with that list evaluated. This procedure is known as FDR thresholding as proposed by Tusher *et al*. (2001).

## 3 RESULTS

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 the same time ('univariate evaluation'). These simulations should reveal differences of the methods in estimating the generalized concordance probability $c'$ under PH and various types of NPH. A second series simulated typical gene expression studies, where we considered a large number of genes with partly correlated expressions competing for selection in the same study ('multivariate evaluation').

### 3.1 Simulation study: univariate evaluation

In this series of simulations, we investigated the effect of the following factors on the distribution of $c'$ estimates in a factorial design, generating 2000 samples of 200 observations for each cell: time-dependency (PH, CH or DH), strength of effects ('small', 'medium' or 'large') and presence and amount of censoring (0, 33, 67%). We generated $\log_2$ gene expression values from a standard normal distribution, and survival time $y$ from a Weibull distribution with shape parameter $a=2$ and scale parameter $b=0.5$. Gene expression was linked to survival time by applying an algorithm of MacKenzie and Abrahamowicz (2002). For time-dependency we considered PH with $\beta(t)=\beta_0$, CH with a time-dependent log HR of $\beta(t)=\beta_0[1+2.88/(1+5t)]$, and DH with $\beta(t)=\beta_0(1+1.86t)$. $\beta_0$ was determined for pre-defined population 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 $\beta_0$ values of log(1.5), log(2) and log(4). Further details of these computations are given in the Supplementary Data, Section 4.1 and Supplementary Figure 5. To simulate censoring we drew a uniformly distributed follow-up time $z$ from U[0, $\tau$] and defined the observed survival time as $t=\min(y,z)$ with status indicator $\mathrm{I}(z>y)$. We determined $\tau$ to obtain proportions of censored times of 33 and 67%.

Figure 2 shows boxplots of the estimates of $c'$ by COX, WHE and CON. The dashed reference lines indicate population values of $c'$. In case of PH all three methods provide approximately unbiased estimates of $c'$, irrespective of the effect size and amount of censoring, and COX shows slight efficiency advantages. In case of NPH (CH, DH) however, WHE and CON have clearly lower bias than COX, which over- or under-estimates depending on the combination of censoring and the type of NPH. With increasing censoring all three methods show variance inflation of similar magnitude. When censoring is combined with time-dependent effects a part of the bias can be attributed to the discrepancy of the population value of $c'$ given follow-up is restricted to a maximum time $\tau$ compared to the unrestricted $c'$. Across all scenarios, the largest observed discrepancy was 0.023.

### 3.2 Simulation study: multivariate evaluation

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 expressions of $p=5000$ features according to a 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 eight 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 the distribution function

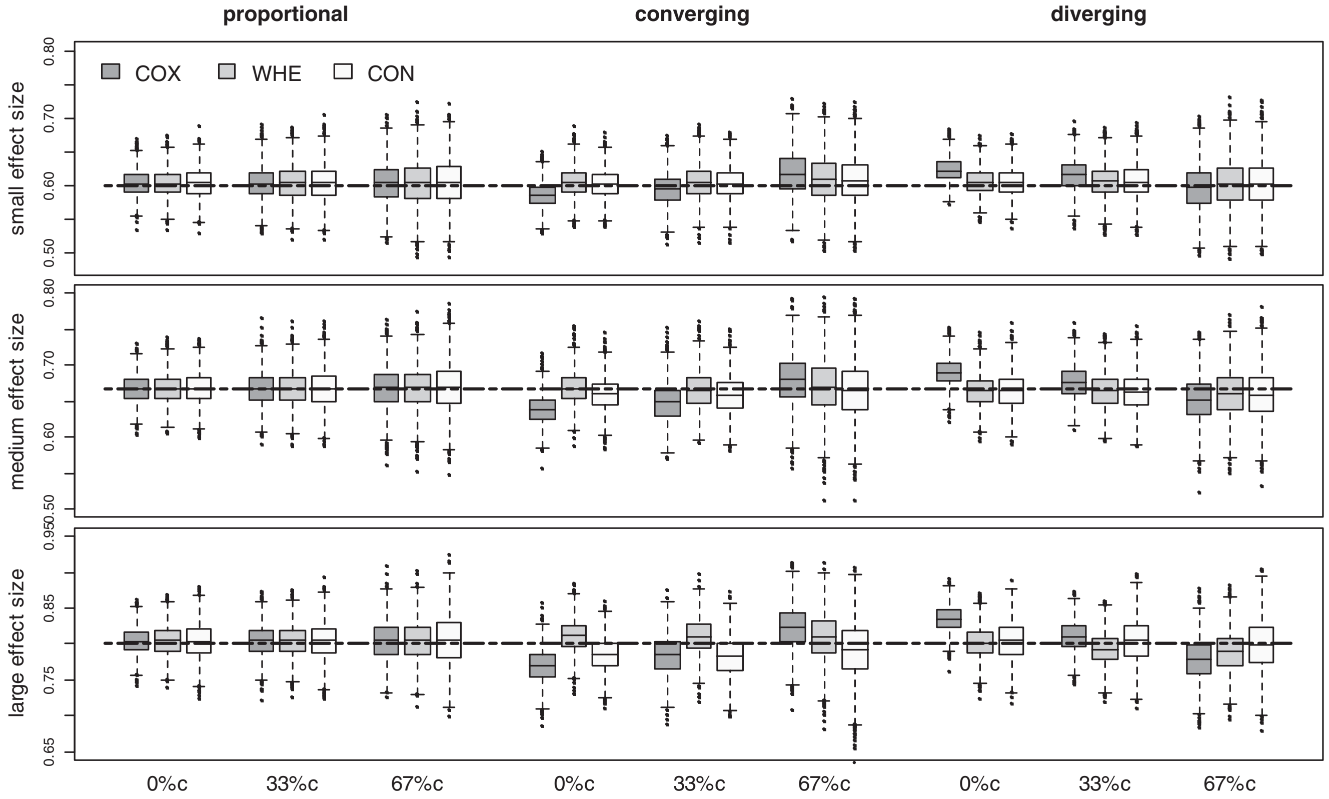

**Fig. 2.** Boxplots of $\hat{c}'$ estimates obtained by Cox (COX), weighted Cox (WHE) and concordance (CON) regression in 2000 simulated data sets with 200 observations each. Dashed lines refer to population values of $c'$. %c, percent censored.

denoted by $F_W(t)$. We linked standard normally distributed gene expressions to survival times, assuming that the hazard of subject $i$ 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 HR of gene $g$ was defined as $\beta_g(t)=\beta_0$ in case of PH, $\beta_g(t)=\beta_0[1+2.88/(1+5t)]$ for CH, and $\beta_g(t)=\beta_0(1+1.86t)$ for DH. The constants $\beta_0$ were set such that average regression effects $\bar{\beta}=\int\beta_g(t)dF_W(t)$ 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.

Each data set was analyzed using COX, WHE and CON and for each gene $\hat{c}'$ was estimated. Genes were ranked by $\hat{c}'_+$ and the $m$ top genes were considered 'selected'. Figure 3 shows some results when $m$ is set to 72, the number of genes truly associated with survival. Results for other choices of $m$ are given in Supplementary Figures 9–14 and Supplementary Tables 3–8.

The average number of correctly selected genes which corresponds to the true positive rate, TPR, under various censoring proportions and sample sizes is graphically compared in Figure 3. The TPR is significantly highest for CON in scenarios with no or 33% censoring (all paired $t$-tests yielded $P<10^{-6}$, cf. Supplementary Table 9). Although CON estimates are on average least biased, their variability is higher than that of WHE or COX, which leads to the impaired performance of CON with 67% censoring. This could be due the unequal weighting of the contributions to the likelihood, which can be severe when there are few 'long survivors' in the data set. To address this issue, we truncated all weights at their 95th percentile and found that the relative loss of efficiency of CON is compensated (CON compared to COX: 31.5 versus 32.0 genes, $P=0.039$ for $N=200$; 41.4 versus 41.7 genes, $P=0.551$ for $N=800$; Table 1). Increasing the sample size from 200 to 800, the number of correctly selected genes increases by ~35% for CON and by ~32% for COX and WHE.

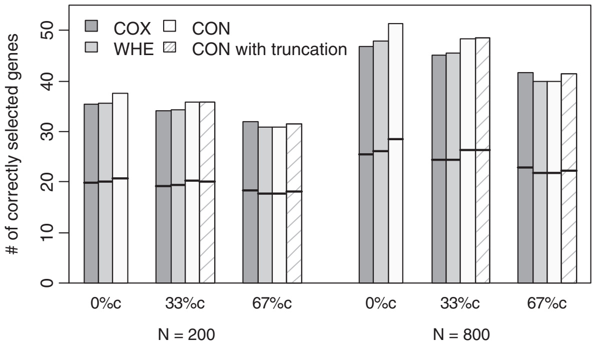

**Fig. 3.** Average number of correctly selected genes by Cox regression (COX), weighted Cox regression (WHE) and concordance regression (CON) without and with truncation of weights from the multivariate evaluation. Lower and upper parts of each bar correspond to correctly selected genes with small and large effect sizes, respectively. %c, percent censored; $N$, sample size.

**Table 1.** Average number of true positive genes in 200 simulated data sets selected by Cox regression/weighted Cox regression/concordance regression with truncation of weights

| | | 0%c | | 33%c | | 67%c | |
|---|---|---|---|---|---|---|---|
| *N* | Hazard | 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[a] | 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[a] | 46.7/47.8/**51.3** | | 45.0/45.5/**48.6** | | 41.7/40.0/41.4 | |

The total number of selected genes was 72 for each method and each scenario. The effect sizes (ES) were set to 'small' for 48 and to 'large' for 24 out of 5000 candidate genes. Total numbers of correctly selected genes were statistically compared between the methods by paired *t*-tests. PH, proportional hazards; DH, diverging hazards; CH, converging hazards; %c, percent censored; ES, effect size; *N*, sample size.

[a] The significantly ($P < 0.01$) highest total number of true positive genes is set in boldface.

If the number of selected genes $m$ is varied, the TPR and similarly the false positive rates change. These changes concern all methods alike, such that we can conclude that the superior performance of CON is independent of a particular choice for $m$. Generally, TPRs are higher with a sample size of 800 compared to 200, but the general conclusions do not change (Supplementary Fig. 7).

In Table 1 and Supplementary Figure 8, we investigate which type of time-dependency is favored by the methods, and whether this depends on censoring and/or sample size. Ideally, genes with equal effect sizes should be selected with equal probability, irrespective of the type of time-dependency (PH, DH or CH), and the censoring pattern. We learn that this ideal situation is best accomplished by CON, which yields the best balance between genes from all types of time-dependency. By contrast, WHE selects CH genes more than others, and with COX the proportions of selected genes of different type change with increasing amount of censoring. These results are independent of the sample size, which affects the number of selected genes in all methods and with all types of genes in the same manner.

### 3.3 Application to real-life studies

We applied univariate COX, WHE and CON to all genes of three microarray data sets and evaluated differences in gene selection. Beer *et al.* (2002) studied the association of survival and gene expression profiles of microarray data of 86 patients with early-stage lung adenocarcinomas. Similarly, Bhattacharjee *et al.* (2001) investigated correlation of gene expression from lung adenocarcinomas with a survival endpoint in 125 patients. In a study by Rosenwald *et al.* (2002) the survival and gene expressions of 240 patients with diffuse large B-cell lymphoma were analyzed. For information regarding pre-processing we refer to our Supplementary Data, Section 3.

In each data set we ranked genes by their estimated absolute effect size $\hat{c}'_+$. 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 $g$ (Storey, 2002). Estimates of $\hat{\pi}_0$ for the three data sets are given in Supplementary Table 1. The $\widehat{\text{FDR}}_{250}$ was then calculated as $\widehat{\text{FDR}}_{250} = \widehat{\text{FP}} \times \hat{\pi}_0/250$.

Results are summarized in Table 2. In all three data sets approximately half of the genes have a negative effect on survival, i.e. $\hat{c}'_g < 0.5$. The range of $\hat{c}'$ varies considerably between the three data sets, with the largest range in the Beer data set (0.254–0.828 computed by CON) and the smallest range in the Bhattacharjee data set (0.412–0.591 computed by CON). The correlation of the absolute effect size estimates $\hat{c}'_+$ from WHE and CON is close to 1 for all data sets, whereas it is considerably smaller when correlating WHE or CON with COX estimates.

If the 250 genes with the largest absolute effect size estimates $\hat{c}'_+$ are selected the best agreement in gene selection is observed between WHE and CON: 71, 79 and 69% in the three data sets. The proportion of genes selected by all three methods is ~50% in all three data sets. The smallest $\widehat{\text{FDR}}_{250}$ values were obtained by COX in the Beer data (0.389), WHE in the Bhattacharjee data (0.841) and CON in the Rosenwald set (0.369). These data dependent advantages of the methods were also observed with higher or lower numbers of selected genes (see Supplementary Table 2). In the Bhattacharjee data the $\widehat{\text{FDR}}_{250}$ estimate from COX selection was larger than 1, a situation which was already anticipated by Tusher *et al.* (2001).

## 4 DISCUSSION

We have introduced $c'$, a semi-parametric generalization of the well-known concordance probability for continuous predictors and

**Table 2.** Results of analysis with Cox (COX), weighted Cox (WHE) and concordance (CON) regression of three data sets

| | No. of genes | No. of obs./number of events | Cor of $\hat{c}'_+$ | | | $\widehat{\mathrm{FDR}}_{250}$ | | | No. of genes selected by | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | COX and WHE | COX and CON | WHE and CON | COX | WHE | CON | COX and WHE | COX and CON | WHE and CON | COX, WHE and CON |
| Beer | 4966 | 86/24 | 0.792 | 0.767 | 0.972 | 0.389 | 0.492 | 0.845 | 167 (67%) | 145 (58%) | 177 (71%) | 134 (54%) |
| Bhattacharjee | 12 600 | 125/71 | 0.829 | 0.829 | 0.994 | 1.053 | 0.841 | 0.923 | 167 (67%) | 158 (63%) | 197 (79%) | 144 (58%) |
| Rosenwald | 7053 | 240/138 | 0.464 | 0.613 | 0.902 | 0.383 | 0.721 | 0.369 | 119 (48%) | 143 (57%) | 172 (69%) | 109 (44%) |

The table summarizes the number of observations and events ('No. of obs/No. of events'), the correlation of the estimated absolute effect size ('Cor of $\hat{c}'_+$'), the estimated false discovery rate ('$\widehat{\mathrm{FDR}}_{250}$') and the number of genes selected in common by COX, WHE and CON when 250 genes are selected based on the estimated absolute effect size. Further results can be found in the Supplementary Data, section 3.

have shown that it is a suitable measure to compare effect sizes between genes irrespective of PHs. In case of PHs, $c'$ can be directly computed from the Cox regression coefficient $\hat{\beta}_C$. In case of NPH, this relationship no longer holds. We have shown by simulation that a novel method, CON, supplies an empirically unbiased estimate of $c'$. WHE as recently proposed approximates this value very well in most but not all cases, and has some efficiency advantages over CON. Cox regression in general fails to provide a consistent estimate of $c'$ in case of NPH, and this bias is further modified by censoring. With large proportions of censored survival times, CON estimates may become inefficient due to unequal weighting of the contributions to the likelihood. As we have demonstrated, truncating weights of CON at a high, e.g. the 95th percentile may then be applied to reduce the variability of the estimates, while the amount of bias introduced is still negligible.

If an association analysis of high-dimensional data with survival involves a gene selection step, then a gene ranking based on estimates of $c'$ may be preferable to a gene ranking based on Cox regression coefficients, because the former does not rely on the assumption of PHs. By contrast, the gene ranking by Cox regression does not only depend on the true effect size of the genes, but also on the realized follow-up distribution. Thus these rankings may not be reproducible under different follow-up schemes. We have shown by analysis of simulated and three real data sets that gene rankings by estimates of $c'$ provided by CON or WHE may yield different results than gene rankings by log HR estimates from Cox regression. The rankings from WHE and CON were more similar than compared to gene rankings from Cox regression, and the agreement of the former two with Cox regression in an absolute sense was low in all three data sets. We have also shown that FDR thresholding based on $c'$ is straightforward.

Our investigation focused on gene selection. This was motivated by our own experience, that most often the microarray platform is mainly used to screen the whole genome for suitable candidate genes. Gene expression data, reduced to a set of, say, 50–400 candidate genes, is then re-determined using a high-sensitivity platform such as real time polymerase chain reaction, and only from these validated expression values statistical models, often on an even further reduced set of genes, will be developed. For some of the methods for prediction of survival from high-dimensional data it was proposed to use gene selection as a first step, e.g. supervised principal components regression (Bair and Tibshirani, 2004; Bair *et al.*, 2006). Application of the methods investigated in this contribution in combined selection and prediction models like the LASSO (Park and Hastie, 2007; Tibshirani, 1997), ridge regression (Verweij and van Houwelingen, 1994) or partial least squares regression (Park *et al.*, 2002) is in principle straightforward. However, our investigation leaves the question open whether the modest improvements in gene selection observed with WHE or CON can contribute to improved estimation of prediction models. Nevertheless, we consider the robustness of these methods to NPH as being of particular advantage for real-life applications.

Thus, one may replace Cox regression by WHE or CON at any stage of model development, to obtain prediction models without having to assume PHs. As final result, such prediction models supply cross-validated risk scores to assess and compare different levels of risk between subjects. An evaluation of the association of the risk scores with survival using $c'$ may improve interpretation as it allows quantifying the impact of the genetic information on survival, provides a direct comparison of subjects and at the same time is robust to violations of the PHs assumption.

## ACKNOWLEDGEMENTS

The authors are grateful to Samo Wakounig, Vienna and Ståle Nygård, Oslo, and three anonymous reviewers for comments that considerably improved the article.

*Funding*: FWF Austrian Science Fund (P18553-N13).

*Conflict of interest*: none declared.

## REFERENCES

Abrahamowicz,M. and MacKenzie,T.A. (2007) Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat. Med.*, **26**, 392–408.

Bair,E. and Tibshirani,R.J. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, 511–522.

Bair,E. *et al.* (2006) Prediction by supervised principal components. *J. Am. Stat. Assoc.*, **101**, 119–137.

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

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.

Bøvelstad,H.M. *et al.* (2007) Predicting survival from microarray data - a comparative study. *Bioinformatics*, **23**, 2080–2087.

Collett,D. (2003) *Modelling Survival Data in Medical Research*. Chapman and Hall, London.

Cox,D.R. (1972) Regression models and life-tables. *J. Royal Stat. Soc. B.*, **34**, 187–220.

Gönen,M. (2007) *Analyzing Receiver Operating Characteristic Curves with SAS*. SAS Institute Inc, Cary, NC.

Gönen,M. and Heller,G. (2005) Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, **92**, 965–970.

Grambsch,P.M. and Therneau,T.M. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515–526.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under an ROCcurve. *Radiology*, **143**, 29–36.

Hess,K.R. (1994) Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Stat. Med.*, **13**, 1045–1062.

Lehr,S. and Schemper,M. (2007) Parsimonious analysis of time-dependent effects in the Cox model. *Stat. Med.*, **26**, 2686–2698.

Lin,D.Y. and Wei,L.J. (1989) The robust inference for the Cox proportional hazards model. *J. Am. Stat. Assoc.*, **84**, 1074–1078.

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

Mann,H.B. and Whitney,D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, **18**, 50–60.

Marubini,E. and Valsecchi,M.G. (1995) *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley, New York.

Moreau,T. *et al*. (1985) A global goodness-of-fit statistic for the proportional hazards model. *Appl. Stat.*, **34**, 212–218.

Ng'andu,N.H. (1997) An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Stat. Med.*, **16**, 611–626.

Park,M.-Y. and Hastie,T. (2007) An L1 regularization-path algorithm for generalized linear models. *J. Royal Stat. Soc. B*, **69**, 659–677.

Park,P.J. *et al*. (2002) Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**, S120–S127.

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.

Schemper,M. *et al*. (2009) The estimation of average hazard ratios by weighted Cox regression. *Stat. Med*., **28**, 2473–2489.

Storey,J. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**, 479–498.

Therneau,T.M. and Grambsch,P.M. (2000) *Modeling Survival Data. Extending the Cox Model*. Springer, New York.

Tibshirani,R.J. (1997) The LASSO method for variable selection in the Cox model. *Stat. Med.*, **16**, 385–395.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Valsecchi,M.G. *et al*. (1996) Evaluation of long-term survival: use of diagnostics and robust estimators with Cox's proportional hazards model. *Stat. Med.*, **15**, 2763–2780.

Verweij,P.J.M. and van Houwelingen,H.C. (1994) Penalized likelihood in Cox regression. *Stat. Med.*, **13**, 2427–2346.

Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometrics*, **1**, 80–83.

Xu,R. and O'Quigley,J. (2000) Estimating average regression effect under non-proportional hazards. *Biostatistics*, **1**, 423–439.

Xu,J. *et al*. (2005) Survival analysis of microarray expression data by transformation models. *Comput. Biol. Chem.*, **29**, 91–94.

# APPENDIX A

## A.1 EQUIVALENCE OF c/(1−c) AND THE HAZARD RATIO UNDER PROPORTIONAL HAZARDS

Assume that $S_j(t)$, $f_j(t)$ and $\lambda_j(t)$ denote the survivor function, the density and the hazard function in group $j$, $j=\{0,1\}$ at time $t$. Under PH we assume that the hazard ratio is constant over time, $\theta(t)=\lambda_1(t)/\lambda_0(t)=\theta$. Thus it follows that $S_1(t)=S_0(t)^{\theta}$. Define as $c=P(T_1<T_0)=\int f_1(t)S_0(t)$. Then

$$c/(1-c)=P(T_1<T_0)/P(T_0<T_1)=\int f_1(t)S_0(t)\bigg/\int f_0(t)S_1(t)$$

$$=\int \theta\lambda_0(t)S_0(t)^{\theta+1}\bigg/\int \lambda_0(t)S_0(t)^{\theta+1}=\theta.$$

Since $\theta=\exp(\beta_C)$, we have $c=\exp(\beta_C)/[1+\exp(\beta_C)]$, thus under PH we can use $\hat{\beta}_C$ from a Cox regression analysis to estimate $c$.