STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:2473–2489
Published online 26 May 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3623

The estimation of average hazard ratios by weighted Cox regression

Michael Schemper*,†, Samo Wakounig and Georg Heinze

Section of Clinical Biometrics, Department for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

SUMMARY

Often the effect of at least one of the prognostic factors in a Cox regression model changes over time, which violates the proportional hazards assumption of this model. As a consequence, the average hazard ratio for such a prognostic factor is under- or overestimated. While there are several methods to appropriately cope with non-proportional hazards, in particular by including parameters for time-dependent effects, weighted estimation in Cox regression is a parsimonious alternative without additional parameters. The methodology, which extends the weighted k-sample logrank tests of the Tarone-Ware scheme to models with multiple, binary and continuous covariates, has been introduced in the nineties of the last century and is further developed and re-evaluated in this contribution. The notion of an average hazard ratio is defined and its connection to the effect size measure $P(X<Y)$ is emphasized. The suggested approach accomplishes estimation of intuitively interpretable average hazard ratios and provides tools for inference. A Monte Carlo study confirms the satisfactory performance. Advantages of the approach are exemplified by comparing standard and weighted analyses of an international lung cancer study. SAS and R programs facilitate application. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: converging hazards; effect size; Prentice test; proportional hazards model; survival analysis; weighted estimation

1. INTRODUCTION

Cox's [1] proportional hazards regression model continues to be the most popular tool in the analysis of censored survival data. Often, the effect of at least one of the prognostic factors included in such a model changes over time which violates the proportional hazards assumption. As a consequence, the average hazard ratio for such a prognostic factor is under- or overestimated.

*Correspondence to: Michael Schemper, Section of Clinical Biometrics, Department for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria.
†E-mail: michael.schemper@meduniwien.ac.at

Contract/grant sponsor: FWF Austrian Science Fund; contract/grant number: P18553-N13

Received 3 October 2008
Accepted 14 April 2009

In order to more appropriately cope with non-proportional hazards the following options are available:

(a) Inclusion of parameters for time-dependent effects.
(b) Changing to a different type of model, e.g. to the proportional odds model or the parametric log-logistic model.
(c) Stratification on a prognostic factor exhibiting non-proportional hazards.
(d) Separate modeling for different time periods.
(e) Weighted estimation and inference for the regression parameters.

There is a large body of literature [2–6] on option (a), which permits very flexible modeling of the time-course of the hazard ratio but also at an increased risk of over-fitting. Nevertheless, this is the approach of choice—in particular, application of fractional polynomials and penalized approaches have been shown [7] to perform satisfactorily—if a more detailed exploration of time dependence is of interest and sample information sufficient. The most frequent departure from proportional hazards is converging hazards. If all prognostic factors exhibit converging hazards (b), the semi-parametric proportional odds model [8–10] and the parametric log-logistic model [11–13] are suitable. These models have optimal power for alternative hypotheses of logistic shift as has the generalized Wilcoxon test for censored data by Prentice [14]. However, it is rare that all prognostic factors of a study have a tendency for converging hazards. In addition, software to fit the proportional odds model is not generally available. Stratification (c) of an analysis by a non-proportional hazards factor is not an attractive option as long as it precludes estimation of its strength and also its test within the fitted model. For a single non-proportional hazards factor, however, Wei and Schaubel [15] recently suggested an estimator for the aggregate factor effect, which is based on the factor-level-specific cumulative baseline hazards estimated under a stratified multiple Cox regression. The fitting of separate Cox models (d) to different time-periods can be very informative but the increased variability of the estimates, the decreased power and increased multiplicity problems make this approach less attractive with small to medium sized samples and/or heavy censoring.

The last of the cited options, (e), is the least known but in certain contexts may be the best choice. It is the topic of this paper, and it permits to obtain a well interpretable average effect if the underlying effect varies in time. Weighted estimation for Cox regression (WCR) was first used by Lin [16] for checking goodness-of-fit of a Cox regression (CR). Schemper [17] demonstrated the suitability of WCR for estimating average hazard ratios when hazards are non-proportional, and Sasieni [18, 19] extensively investigated statistical properties of WCR. He concluded [18] that the method is '... a useful addition to the standard Cox regression estimator ...'. Xu and O'Quigley [20] showed that in the absence of censoring the usual, unweighted partial likelihood estimate can be interpreted as an average regression effect, independent of the existence of proportional hazards, and they use weighted score equations to preserve this property under censoring. The notion of an average regression effect differs from the notion of the average hazard ratio of this paper insofar as the latter takes into account the decreasing number of individuals affected by the hazards with increasing follow-up time.

Based on an experience of 15 years with applying WCR to clinical studies of survival, we now review the methodology including recent developments by our group, and re-evaluate its role for modeling survival data. In the sequel we define the notion of an average hazard ratio (in Section 2), and revisit its estimation and corresponding inference by Cox regression (in Section 3). We empirically compare different definitions of (average) hazard ratios on the population

DOI: 10.1002/sim

level and quantify advantages and disadvantages of weighted estimation by simulation experiments (in Section 4). Comparative analyses of a lung cancer study are presented in Section 5. Section 6 summarizes indications for WCR, and introduces computer programs in SAS and R which permit convenient application.

2. THE AVERAGE HAZARD RATIO

Hazards as considered by survival analysis often will be expected to change over time, and hazard ratios similarly. Therefore, statistical methods either require specification of this time dependence or they consider a value averaged over time. In this paper we only deal with the latter and now introduce definitions of average hazard ratios which keep their interpretability also under non-proportional hazards. For the ease of presentation we start with the two-sample case without censoring, but keep the final generalizability to a multiple regression model for censored survival times in mind.

We consider the following three definitions of average hazard ratios:

$$s\text{AHR} = \int \frac{h_1(t)}{h_0(t)} w(t) f(t)\,\mathrm{d}t \tag{1}$$

$$g\text{AHR} = \exp\left[\int \log\left(\frac{h_1(t)}{h_0(t)}\right) w(t) f(t)\,\mathrm{d}t\right] \tag{2}$$

$$\text{AHR} = \frac{\int (h_1(t)/h(t)) w(t) f(t)\,\mathrm{d}t}{\int (h_0(t)/h(t)) w(t) f(t)\,\mathrm{d}t} \tag{3}$$

where sAHR stands for 'simple average hazard ratio', gAHR for 'geometric average hazard ratio' and AHR is an average hazard ratio definition by Kalbfleisch and Prentice [21] which is our preferred definition due to reasons given below. In definitions (1)–(3) $h_0(t)$ and $h_1(t)$ denote the hazards of groups G_0 and G_1 at time t, respectively, and $h(t) = h_0(t) + h_1(t)$. The weight function $w(t)$ is chosen to reflect the relative importance attached to the hazard ratios in different time periods, the most basic choices being $w(t) = 1$ and $w(t) = S(t)$, the survival function, or equivalently, the proportion of individuals affected by a hazard ratio at t. In the sequel the subscripts Cox and WCox will be used to denote the specific choices of $w(t) = 1$ and $w(t) = S(t)$ for definitions (1)–(3), e.g. AHR_{WCox} symbolizes definition (3) with $w(t) = S(t)$. In the population definitions (1)–(3) the density of the events in time (which result in contributions to the partial likelihood in Cox's model) is needed and symbolized by $f(t)$.

While sAHR is very intuitive it is not symmetric in $h_0(t)$ and $h_1(t)$ and therefore not favored. Definition gAHR does not suffer from this shortcoming but represents a geometric average hazard ratio. Definition AHR defines each hazard relative to the sum of hazards and nicely connects to other measures of interest as dealt with below.

The definitions of AHRs (1)–(3) can be modified for continuous covariates x, e.g. AHR then becomes

$$\text{AHR} = \frac{\int [\int (h_{x+\Delta}(t)/h(t)) f(t) w(t)\,\mathrm{d}t] g(x)\,\mathrm{d}x}{\int [\int (h_x(t)/h(t)) f(t) w(t)\,\mathrm{d}t] g(x)\,\mathrm{d}x} \tag{4}$$

DOI: 10.1002/sim

where Δ denotes the change in x for which the hazard ratio is reported, $h(t)=h_x(t)+h_{x+\Delta}(t)$, and $g(x)$ the density of x.

While we do not present the formalism for AHRs with multiple covariates here, corresponding estimates are indeed made available by the WCR approach and will be dealt with in succeeding sections.

We now turn to the interesting issue of interpretation of AHR under the two basic weight functions: the choice $w(t)=1$ leads to a hazard ratio of standard Cox type, which is an average of the hazard ratios at all death times, also under non-proportional hazards. Grambsch and Therneau [22] have shown that the log hazard ratio at time t can be represented by an average log hazard ratio plus the expected value of the scaled Schoenfeld residual at t. Since the sum of these residuals, with one residual per death, is zero, the Cox hazard ratio can be interpreted as an average over the observed death times. Thus, all times are taken as equally important, regardless of the different numbers of individuals at risk. The choice of $w(t)=S(t)$ is preferable if the relative importance of hazard ratios at different times is considered as being proportional to the numbers of individuals at risk at these times.

Support for the choice of $w(t)=S(t)$, more precisely for $\mathrm{AHR}_{\mathrm{WCox}}$, also comes from an interesting (approximate) connection with another important statistic, as dealt with subsequently.

The probability $P(T_1<T_0)$ 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 is an intuitive and elementary non-parametric measure of effect size, characterizing the degree of separation of the distributions of the survival times of two groups. For uncensored T_0 and T_1 it is directly related to the statistics of the non-parametric two-sample tests by Wilcoxon [23] and by Mann and Whitney [24] and it is identical to the area under an ROC curve (cf. Hanley and McNeil [25]) as well as to the c-index (cf. Harrell [26, Section 5.5]). Therefore, we will term $P(T_1<T_0)$ *concordance probability*. Other names have been proversion probability, probabilistic index, stochastic superiority $P(Y>X)$, and individual exceedence probability. This effect size measure is at the origin of non-parametrics and has received much attention in the past: more recently also a special issue of this journal was devoted to it (volume 25, issue 4). Of particular interest here is the relationship of this measure to the odds-of-concordance, OC, which Begun and Reid [27], in the context of proportional hazards, cited as a hazard ratio:

$$\mathrm{OC}=\frac{P(T_1<T_0)}{1-P(T_1<T_0)} \tag{5}$$

or, equivalently,

$$P(T_1<T_0)=\frac{\mathrm{OC}}{\mathrm{OC}+1} \tag{6}$$

$P(T_1<T_0)$ as well as OC is based on all equally weighted pairwise comparisons of the survival times between groups G_0 and G_1, and evidently does not require proportional hazards to be interpretable. It is shown in Appendix A that OC is identical to AHR with the choice of $w(t)=(S_0(t)f_1(t)+S_1(t)f_0(t))/(f_0(t)+f_1(t))$, a weighted average of the survival functions for G_0 and G_1. For this weight function AHR simplifies to

$$\mathrm{AHR}_{\mathrm{OC}}=\frac{\int h_1(t)S_0(t)S_1(t)\,\mathrm{d}t}{\int h_0(t)S_0(t)S_1(t)\,\mathrm{d}t} \tag{7}$$

DOI: 10.1002/sim

Use of the appealing OC within a multiple Cox model would result in a cumbersome fitting process—due to the more involved definition of the weight. Therefore, it is of interest whether $\mathrm{AHR}_{\mathrm{WCox}}$ and $\mathrm{AHR}_{\mathrm{OC}}$ (=OC) are sufficiently close for the typical range of hazard ratios encountered in practice. This will be explored below.

We considered five typical situations described by panels A–E of Figure 1 for an empirical comparison of OC (=$\mathrm{AHR}_{\mathrm{OC}}$) and of $\mathrm{AHR}_{\mathrm{WCox}}$. Furthermore, we provide results for $\mathrm{AHR}_{\mathrm{Cox}}$, $g\mathrm{AHR}_{\mathrm{WCox}}$ and $g\mathrm{AHR}_{\mathrm{Cox}}$. The situations were selected in such a way that OC is 2 for A–C and 1 for D and E: (A) proportional hazards for groups G_0 and G_1, (B) non-proportional, converging hazards with dominating early differences in the survival functions of G_0 and G_1, (C) non-proportional, diverging hazards with dominating late differences, (D) identical hazards, and (E) crossing hazards and survival functions representing a non-proportional hazards null hypothesis in terms of OC=1 or $P(T_0<T_1)=0.5$.

Table I gives the results: it is obvious that under proportional hazards all definitions of average hazard ratios assume identical values because $h_1(t)/h(t)$ in these definitions is constant for all t and any choice for $w(t)$ does not affect AHR which becomes h_1/h_0. Under non-proportionality $\mathrm{AHR}_{\mathrm{OC}}$

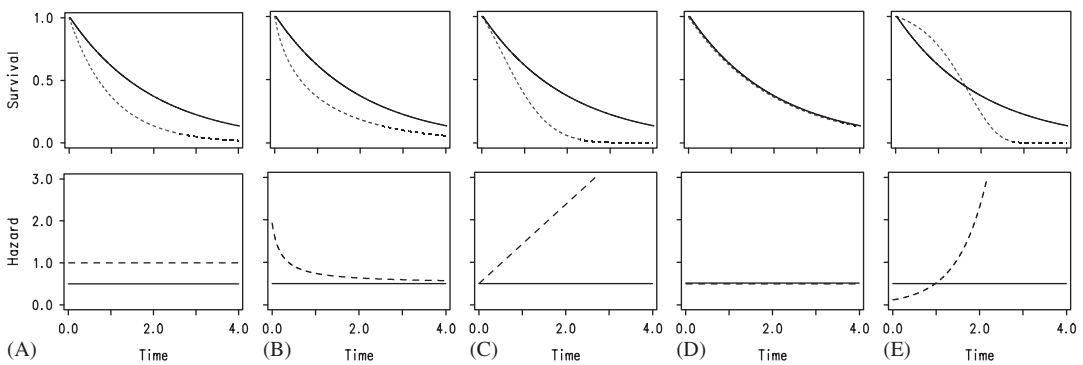

Figure 1. Survival and hazard functions of two groups G_0 and G_1 for five typical situations of group effects. The survival of group G_0 is characterized by a constant hazard $h_0(t)=0.5$. The hazard functions for group G_1 are defined by (A) $h_1(t)=2h_0(t)$ ('proportional hazards'), by (B) $h_1(t)=(1+2.88/(1+5t))\cdot h_0(t)$ ('converging hazards'), by (C) $h_1(t)=(1+1.86t)\cdot h_0(t)$ ('diverging hazards'), by (D) $h_1(t)=h_0(t)$ ('identical hazards'), and by (E) $h_1(t)=0.11\exp(1.5t)$ ('crossing hazards').

Table I. Hazard ratios for situations A–E, defined in Figure 1, according to OC, $g\mathrm{AHR}_{\mathrm{WCox}}$, $\mathrm{AHR}_{\mathrm{WCox}}$, $g\mathrm{AHR}_{\mathrm{Cox}}$ and $\mathrm{AHR}_{\mathrm{Cox}}$.

Definitions of hazard ratios	A	B	C	D	E
OC (=$\mathrm{AHR}_{\mathrm{OC}}$)	2.00	2.00	2.00	1.00	1.00
$g\mathrm{AHR}_{\mathrm{WCox}}$	2.00	2.02	2.10	1.00	1.08
$\mathrm{AHR}_{\mathrm{WCox}}$	2.00	1.98	2.03	1.00	1.02
$g\mathrm{AHR}_{\mathrm{Cox}}$	2.00	1.66	2.98	1.00	3.14
$\mathrm{AHR}_{\mathrm{Cox}}$	2.00	1.63	2.73	1.00	1.82

DOI: 10.1002/sim

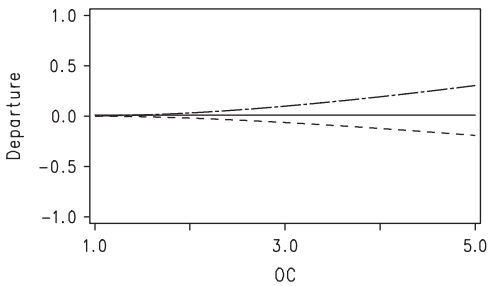

Figure 2. Departure of the hazard ratio AHR_{WCox} from the one defined by $AHR_{OC}(=OC)$. At $OC=2$ the AHR_{WCox} values are those for situations A, B and C of Figure 1. $h_1(t)$ was multiplied by a constant required to obtain OC values in the range 1 to 5. Results for the A-, B- and C-type situations are symbolized by '——', '– – –' and '–·–·–', respectively.

and AHR_{WCox} lead to almost identical values, $gAHR_{WCox}$ values being quite similar. Values of $gAHR_{Cox}$ and AHR_{Cox} substantially differ from those of AHR_{OC}, $gAHR_{WCox}$ or AHR_{WCox}.

Though a hazard ratio of 2 in clinical studies of survival often is considered a strong effect we additionally explored differences of AHR_{OC} and AHR_{WCox} in the presence of hazard ratios up to 5 according to AHR_{OC}. The strong agreement of both measures for typical situations is confirmed by Figure 2 and indicates that we could consider AHR_{WCox}, which more easily lends itself to implementation in a multiple Cox model, as a close surrogate to AHR_{OC} or OC, at least for hazard ratios up to 4 (where the relative discrepancy is between 0 and 5 per cent, depending on the degree of nonproportionality).

Concluding on the relative merits for interpreting average hazard ratios based on either $w(t)=1$ or $w(t)=S(t)$, we clearly prefer the latter under non-proportionality, last but not the least because $P(T_1<T_0)$ (as can be obtained by using (6)) always gives a clear-cut answer of which level of a prognostic factor is more favorable. We have seen that under proportionality both weight functions lead to identical (average) hazard ratios.

In general neither $gAHR_{Cox}$, AHR_{Cox}, $gAHR_{WCox}$, AHR_{WCox} nor AHR_{OC} are population values estimated by the CR and WCR models. However, the asymptotic values for the CR and WCR estimators under non-proportional hazards lack intuitive simplicity (see, e.g. Struthers and Kalbfleisch [28, equation (2.5)]). Therefore, we find it more convincing to show the degree to which the above, conceptually simple and plausible definitions are approximated by the estimation in the CR and WCR models. This will be dealt with by Section 4.

3. ESTIMATION IN COX REGRESSION

3.1. Weighted estimation within Cox's model

In a sample of n individuals, we observe m distinct and uncensored survival times t_j $(1\leqslant j\leqslant m)$ among the n possibly censored survival times t_i $(1\leqslant i\leqslant n)$. A covariate vector $x_i=(x_{i1},\ldots,x_{ir},\ldots,x_{ik})$ is related to each individual as is a censoring indicator η_i (1 for censored, 0 for dead). The set of individuals alive and uncensored prior to t_j, the risk set, is denoted by R_j, as is also

DOI: 10.1002/sim

the size of this set. A vector β of k regression parameters is to be estimated. Then the log partial likelihood for Cox's [1] model is defined as

$$\log L(\beta)=\sum_{j=1}^{m}\left[x_j\beta-\log\left\{\sum_{h\in R_j}\exp(x_h\beta)\right\}\right]=\sum_{j=1}^{m}l_j \tag{8}$$

where l_j is the contribution to the log likelihood at failure time t_j. Setting the first derivatives of $\log L(\beta)$,

$$\frac{\partial\log L(\beta)}{\partial\beta_r}=\sum_{j=1}^{m}\left[x_{jr}-\frac{\sum_{h\in R_j}x_{hr}\exp(x_h\beta)}{\sum_{h\in R_j}\exp(x_h\beta)}\right]=\sum_{j=1}^{m}\frac{\partial l_j}{\partial\beta_r},\quad 1\leqslant r\leqslant k \tag{9}$$

to zero yields estimates of β which have to be obtained by iteration, usually using the Newton–Raphson technique.

The elements of the corresponding information matrix I are

$$I_{rs}=\frac{-\partial^2\log L(\beta)}{\partial\beta_r\partial\beta_s}=\sum_{j=1}^{m}\frac{-\partial^2 l_j}{\partial\beta_r\partial\beta_s}=\sum_{j=1}^{m}\left\{\frac{\sum_{h\in R_j}x_{hr}x_{hs}\exp(x_h\beta)}{\sum_{h\in R_j}\exp(x_h\beta)}\right.$$
$$\left.-\frac{[\sum_{h\in R_j}x_{hr}\exp(x_h\beta)][\sum_{h\in R_j}x_{hs}\exp(x_h\beta)]}{[\sum_{h\in R_j}\exp(x_h\beta)]^2}\right\}\quad(1\leqslant r,s\leqslant k) \tag{10}$$

We now introduce weight functions $w(t)$ which can be chosen to weight the contributions to the log partial likelihood at the m uncensored survival times differently. Weighted maximum likelihood estimates $\hat{\beta}_r$ of regression parameters β_r, $1\leqslant r\leqslant k$, are derived as solutions to weighted score equations

$$\sum_{j=1}^{m}w(t_j)\frac{\partial l_j}{\partial\beta_r}=0 \tag{11}$$

The choice of weights $w(t)$ will be dealt with in Section 3.3.

The asymptotic properties of the weighted estimator have been dealt with by Lin [16] and Sasieni [18] under proportional hazards. While an exhaustive treatment under non-proportionality of hazards along the lines of Struthers and Kalbfleisch [28] is beyond the scope of this contribution, the Monte Carlo results of Section 4 and those provided in a technical report [29] confirm satisfactory empirical performance.

3.2. Inference for weighted estimates

We consider three approaches to arrive at a covariance matrix for WCR estimates: (i) is via a sandwich estimate used by Lin [16] and Sasieni [18], (ii) is based on robust covariance matrices as introduced by Lin and Wei [30], and (iii) is based on the jackknife:

(i) Lin [16] and Sasieni [18] propose a covariance matrix of $\hat{\beta}$, $V_{LS}=A^{-1}BA^{-1}$, where the elements of A and B are

$$A_{rs}=\sum_{j=1}^{m}w(t_j)\frac{-\partial^2 l_j}{\partial\beta_r\partial\beta_s} \tag{12}$$

DOI: 10.1002/sim

and

$$B_{rs}=\sum_{j=1}^{m} w(t_j)^2 \frac{-\partial^2 l_j}{\partial\beta_r \partial\beta_s} \quad (1\leqslant r,s\leqslant k) \tag{13}$$

respectively, the term $-\partial^2 l_j/\partial\beta_r\partial\beta_s$ has been specified by (10). Thus, the scaling of the covariance matrix agrees with the scaling of the estimates. For $w(t_j)=1$, V_{LS} reduces to I^{-1} of CR.

(ii) An alternative definition of the covariance matrix (cf. Therneau and Grambsch [31, Chapter 7]) is $V_{LW}=A^{-1}(U'U)A^{-1}$, where A is defined by (12) and U is the matrix of score residuals for WCR:

$$U_{ir}=(1-\eta_i)w(t_i)\left\{x_{ir}-\frac{\sum_{h\in R_i} x_{hr}\exp(x_h\beta)}{\sum_{h\in R_i}\exp(x_h\beta)}\right\} - \sum_{h:t_h\leqslant t_i}(1-\eta_h)w(t_h)\frac{\exp(x_i\beta)}{\sum_{l\in R_h}\exp(x_l\beta)}\left\{x_{ir}-\frac{\sum_{l\in R_h} x_{lr}\exp(x_l\beta)}{\sum_{l\in R_h}\exp(x_l\beta)}\right\} \tag{14}$$

(iii) Consider the jackknife values $J_i=\hat\beta-\hat\beta_{(i)}$, $(1\leqslant i\leqslant n)$, where $\hat\beta_{(i)}$ is the result of a fit that includes all individuals except the ith. The J_i are now the row vectors of matrix J which is of dimension $n\times k$. Then the jackknife estimate of the covariance matrix of $\hat\beta$, V_J, is given by $V_J=((n-1)/n)(J-\bar J)'(J-\bar J)$, where $\bar J$ is the matrix of column means of J (i.e. from each entry of J the corresponding average over individuals is subtracted).

The robust estimate (ii) used here is algebraically identical (cf. Therneau and Grambsch [31, pp. 159–160]) to Lin and Wei's [30] sandwich estimator of the covariance matrix, which they have shown to be robust to several misspecifications of the Cox model, in particular to the lack of proportional hazards.

An empirical comparison of the estimators V_{LS}, V_J and V_{LW} is provided in Section 4. Based on either of these covariance matrices, Wald-type tests and confidence intervals are available.

3.3. Choice of weights

The function $w(t)$ has been introduced in the definition of an average hazard ratio (3) to differently weight, at different t, the contributions of $h_1(t)/h(t)$ (and equivalently of $h_0(t)/h(t)$) to an AHR. It was noted in Section 2 that the choice of $w(t)=1$ leads to hazard ratios of standard Cox type, while the choice of $w(t)=S(t)$ produces average hazard ratios $\mathrm{AHR}_{\mathrm{WCox}}$.

Analogously, for estimation and inference (11)–(14) with uncensored samples $w(t)=1$ and $w(t)=\hat S(t)$ lead to a CR and a WCR, respectively. $\hat S(t)$ denotes the Kaplan–Meier estimator of the survival function and reflects the relative importance attributed to the log hazard ratios at different times t, i.e. it results in a weighting by the expected number of individuals affected by the hazards at t.

With censored samples, $w(t)=\hat S(t)\hat G(t)^{-1}$ is required for a WCR. $\hat G(t)$ denotes the Kaplan–Meier estimator of the censoring or potential follow-up distribution, estimated like $\hat S(t)$ but with the meaning of the censoring indicator η reversed. $\hat G(t)^{-1}$ is used to compensate the attenuation in observed events due to the earlier censorship. This inverse probability weighting permits reconstructing $f(t)$ from a censored sample—in the time range covered, i.e. till the last event. As in the

DOI: 10.1002/sim

presence of time-dependent effects, the magnitude of contributions to the log partial likelihood (8) differs for different t, and as censoring eliminates such contributions with increasing probability for progressing t, the 'reconstruction' of an uncensored situation by means of $\hat{G}(t)^{-1}$ is required for estimates of average hazard ratios under censoring.

In addition, for a CR under censoring $w(t)=1$ needs to be replaced by $w(t)=\hat{G}(t)^{-1}$ in order to preserve the interpretation of an average regression effect [20], i.e. an expected average with respect to survival times, had censoring not occurred. Under proportional hazards, however, both weight functions will lead to identical estimates for the CR.

One final word of caution on the use of $\hat{G}(t)^{-1}$: if relatively few individuals in a sample experience much longer follow-up than the rest of the sample, the weight at these individuals' definite survival times becomes large, which will reduce the stability of results. Therefore, in such cases it may be preferable to limit the time range for analysis to a more compact follow-up, rather than by the last event.

As with uncensored samples $\hat{G}(t)=1$ for any t, the weighting then simplifies to $w(t)=\hat{S}(t)$. In this case, and with a single binary covariate, the corresponding score test becomes a conditional permutation analogue to the fully permutational Mann–Whitney U-test (cf. [32]). It can be concluded that its property of optimal power against shift between two logistic distributions (cf., e.g. [33]) extends to the WCR model using $\hat{S}(t)$ weights. Even if we do not expect precise logistic shift alternatives, WCR using $\hat{S}(t)$ weights will be more powerful than CR under converging hazards when the effect of initial prognostic factors such as treatment disappears with time. Conversely, under proportional hazards, the weighting used in CR leads to optimal power.

In an analogous way as the score test in a CR for a single binary covariate reduces to Mantel's [34] test, Prentice's [14] test can be represented as a score test in WCR using $\hat{S}(t)$ weights. In fact, the methodology presented in this section allows to generalize all k-sample weighted logrank tests of the Tarone and Ware [35] scheme to models with multiple, binary and continuous covariates. In particular, it generalizes the tests by Breslow [36], Tarone and Ware [35], Prentice [14] and Harrington and Fleming [37] by setting $w(t)=R(t)$, $w(t)=\sqrt{R(t)}$, $w(t)=\hat{S}(t)$ and $w(t)=[\hat{S}(t)]^{\rho}$, respectively, $R(t)$ denoting the number of individuals at risk of death at time t and $\rho \geqslant 0$.

In this paper, however, we do not investigate these score tests within WCR. In our thinking, the most interesting property of weighting functions $w(t)=\hat{S}(t)\hat{G}(t)^{-1}$ is the very intuitive interpretability of the resulting average hazard ratio $\exp(\hat{\beta})$, independent of proportional hazards. Because of our emphasis on this property we also do not deal with further choices for $w(t)$ which might be of interest in other contexts.

4. EMPIRICAL STUDIES

The empirical performance of CR and WCR was explored by a comprehensive Monte Carlo study. The study uses the five situations introduced in Section 2 and graphically represented by Figure 1. In all investigations reported below, survival times were randomly drawn from the distributions defined in the legend to Figure 1.

DOI: 10.1002/sim

4.1. Performance of different estimators of the variance

We compare the performance of the three alternative estimators of the variance, more precisely of the standard error of a single (e.g. treatment) parameter, with the actual empirical standard error. From among various combinations of sample size and balance of a dichotomous covariate we present here WCR and CR results for $n=40$, balance 1:1, and for $n=80$, balance 1:4. While results for further combinations as well as after inclusion of a second covariate are contained in a technical report [29], typical performance can readily be studied from the results in Table II. The standard errors given by the entries have been rounded to 2 decimal digits and multiplied by 100, their variability, in terms of standard errors of the standard error estimates have been rounded to 3 decimal digits and multiplied by 1000. All entries are based on 10 000 simulated samples and thus are virtually unaffected by randomness.

From Table II we conclude that neither results by V_{LW} nor by V_{LS} (which under unweighted estimation is identical to the CR variance) perfectly agree with the actual empirical standard error, but departures are mild. The jackknife, once termed a 'quick and dirty' method, has the smallest bias, but computationally it is the slowest alternative. In terms of the variability of these estimates, V_{LS} performs best and V_J worst. Thus, V_{LW} seems to be a good compromise between bias and efficiency, and is used in the remainder of this paper. Results based on the other estimators, available in the technical report [29], are more similar than would be expected from comparing the standard errors of Table II.

4.2. Comparisons of parameter estimates by WCR and CR, and of the power and size of tests

Comparing the median values of the hazard ratio estimates from 10 000 samples by WCR and CR in Table III with the corresponding population values AHR_{WCox} and AHR_{Cox} in Table I, we

Table II. Performance of different estimators of the standard error for weighted and unweighted estimates in uncensored samples.

	$n=40$, bal$=1:1$			$n=80$, bal$=1:4$		
	A	B	E	A	B	E
WCR						
Empirical	39	38	41	36	32	44
V_{LS}	38 (16)	37 (12)	38 (10)	36 (24)	36 (20)	32 (20)
V_{LW}	37 (20)	36 (15)	37 (20)	35 (35)	31 (32)	40 (33)
V_J	39 (27)	39 (24)	38 (22)	37 (39)	33 (36)	43 (35)
CR						
Empirical	36	37	36	31	27	43
V_{LS}	34 (23)	33 (17)	36 (20)	30 (18)	29 (10)	33 (27)
V_{LW}	33 (24)	33 (20)	35 (25)	30 (36)	26 (34)	41 (39)
V_J	36 (62)	37 (68)	36 (28)	33 (43)	29 (45)	45 (48)

Note: Each entry is based on 10 000 simulated samples; n and bal denote total sample size and degree of balance of a dichotomous covariate, respectively.
The letters A, B and E refer to the situations of underlying survival and hazard functions, specified in Figure 1. The estimates of the standard error by V_{LS} [16, 18], V_{LW} [30] and V_J (jackknife), are contrasted with the observed standard error of the simulated samples ('empirical'). Numbers in parentheses show the variability of the individual estimates of standard errors.

DOI: 10.1002/sim

Table III. Median values of hazard ratio estimates, and power (size) of two-sided tests ($\alpha = 0.05$), all for weighted and unweighted estimates using the robust variance estimates V_{LW}, in uncensored samples.

	$n=40$, bal $=1:1$					$n=80$, bal $=1:4$				
	A	B	C	D	E	A	B	C	D	E
Median values of hazard ratio estimates										
WCR	2.02	1.99	2.06	1.00	1.01	2.02	2.00	2.06	1.00	1.00
CR	2.03	1.71	2.51	1.00	1.56	2.00	1.62	2.73	0.99	1.85
Power (*size*)										
WCR	47	48	46	6	7	52	60	45	6	7
CR	57	37	77	6	23	65	48	82	7	33

Note: Each entry is based on 10 000 simulated samples; n and bal denote total sample size and degree of balance of a dichotomous covariate, respectively.
The letters A–E refer to the situations of underlying survival and hazard functions, specified in Figure 1.

learn that the latter are reasonably well estimated by WCR and CR, in particular with $n=80$. Furthermore, WCR estimates the underlying OC values (2.00 for situations A–C and 1.00 for D and E) quite well. Under non-proportionality (situations B, C and E) median values of estimated hazard ratios differ substantially for CR and for WCR.

Results on power in Table III confirm that with proportional hazards alternatives (situation A) CR estimation is more powerful than the WCR approach. With the non-proportional hazards situations WCR is more powerful under the dominating early effects of B while the reverse is true under the dominating late effects of C. However, in either case the interpretability of estimated hazard ratios is more intuitive under WCR.

Under situation E, for which $H_0: \mathrm{OC}=1$ holds, tests within WCR approximately keep their size, while the tests within CR reject this null hypothesis with a power of 23 and 33 per cent for $n=40$ and $n=80$, respectively.

4.3. Results on estimation from censored samples

In the investigations of this subsection, the survival times generated under situations A–E were subjected to administrative censoring using the model of a medical study. Individuals were assumed to enter the study at a constant rate in the interval $(0, \tau)$ and then to die according to the prescribed survival distribution. For each combination of situation (A–E) and expected percentage of censored survival times (0, 33, 67) a value of τ, the time of analysis, was determined to achieve this percentage of censoring. In Table IV population values of average hazard ratios ($\mathrm{AHR_{OC}}$) are contrasted with estimates obtained under weighting by $\hat{S}(t)\hat{G}(t)^{-1}$ and by $\hat{S}(t)$ only. Note that in the non-proportional hazards situations B, C and E the population values are slightly affected by increasing censoring (visibly only for 67 per cent censored): if τ moves closer to the origin, the time range for which an average of the time-dependent hazard ratio is obtained changes as well. From Table IV we learn that already without censoring there is a small departure of the estimates from the population values for $\mathrm{AHR_{OC}}$, because WCR relates to the simpler $\mathrm{AHR_{WCox}}$. (See Table I for a comparison of $\mathrm{AHR_{OC}}$ and $\mathrm{AHR_{WCox}}$.) Under censoring, estimation using weights $\hat{S}(t)\hat{G}(t)^{-1}$ is relatively stable, in contrast to weighting by $\hat{S}(t)$ only.

DOI: 10.1002/sim

Table IV. OC population values versus estimates by AHR_{WCox} under censoring, using different weights.

Per cent censored	Method	A	B	C	D	E
0	*True value*	2.00	2.00	2.00	1.00	1.00
	$S(t)$	2.02	1.98	2.09	1.01	1.02
	$S(t)G(t)^{-1}$	2.02	1.98	2.09	1.01	1.02
33	*True value*	2.00	1.99	2.02	1.00	1.01
	$S(t)$	2.03	2.10	1.96	1.01	0.88
	$S(t)G(t)^{-1}$	2.03	1.98	2.08	1.01	1.01
67	*True value*	2.00	2.24	1.82	1.00	0.88
	$S(t)$	2.03	2.50	1.64	1.02	0.62
	$S(t)G(t)^{-1}$	2.02	2.26	1.90	0.99	0.96

Note: *True value* denotes the population values for AHR_{OC} (=OC) obtained by integration, $S(t)$ and $S(t)G(t)^{-1}$ denote weighted estimation using $w(t)=\hat{S}(t)$ and $w(t)=\hat{S}(t)\hat{G}(t)^{-1}$, respectively, and the letters A–E refer to the situations of underlying survival and hazard functions, specified in Figure 1. The sample size for the estimates is 10 000.

5. EXAMPLE

The International Adjuvant Lung Cancer Trial (IALT) was designed to evaluate the effect of cisplatin-based adjuvant chemotherapy on survival after complete resection of non-small-cell lung cancer. Between 1995 and 2000 a total of 1867 patients were recruited and randomized to either the chemotherapy or the control arm. Statistical results of this large multi-center study of survival, based on follow-up till 2002, have been provided by 'The IALT Collaborative Group' [38]. Our reanalysis of this study is confined to the data from two major centers. The analysis is based on the survival times of 199 patients (30 per cent censored), treatment (chemo vs control) and the stratification factors used in the randomization procedure: type of surgery (pneumonectomy vs other), pathological stage of disease (I vs II vs III) and center (0 vs 1). The underlined levels of these factors were taken as reference categories in regression models.

Effects of the prognostic factors on survival were analyzed by simple and multiple CR and WCR models. Departures from a main effects CR model were explored by means of additional terms for pairwise interactions, for a non-linear effect of stage, and for (log-)time dependencies of main effects. All but one of the p-values from these exploratory tests exceeded a 5 per cent threshold. The center effect exhibited time dependence ($p<0.0001$) as is also demonstrated by the unadjusted survival functions and a Schoenfeld residual plot of Figure 3. Adjusted and unadjusted results of an analysis of the four main effects obtained by unweighted and weighted estimation (using $w(t)=\hat{S}(t)\hat{G}(t)^{-1}$) are presented in Table V. We learn that it does not make much difference whether WCR or CR is employed for the proportional factors; however, for the non-proportional center effect it makes much difference, in terms of the hazard ratio as well as in terms of the associated p-value. Interpretation of the center results from the weighted approach is more appealing to us. The variation explained due to Schemper and Henderson [39] increases from 1.5 to 2.9 per cent for the unadjusted center effect if we move from unweighted to weighted estimation.

In the left panel of Figure 4 survival functions based on CR and WCR for the static parameter for center are superimposed on Kaplan–Meier estimates. We recognize that the survival functions

DOI: 10.1002/sim

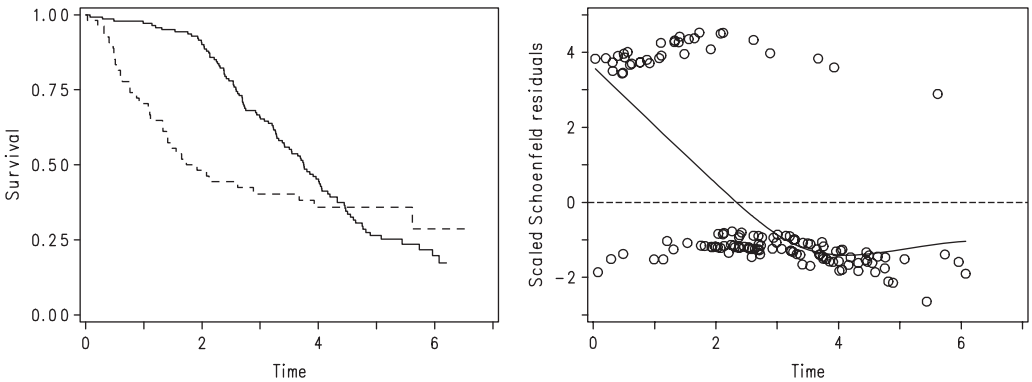

Figure 3. IALT study: Kaplan–Meier estimates of survival functions for centers 0 '——' and 1 '– – –' (left panel); plots of Schoenfeld residuals for centers (right panel).

Table V. Hazard ratios (p-values) for the prognostic factors of the IALT study subgroup, according to simple/multiple Cox regressions using unweighted/weighted estimation.

	Unadjusted effects		Adjusted effects	
Estimation	Unweighted	Weighted	Unweighted	Weighted
Factors				
Treatment	0.88 (0.44)	0.89 (0.50)	0.79 (0.19)	0.82 (0.30)
Stage	1.74 (<0.001)	1.74 (<0.001)	1.95 (<0.001)	1.95 (<0.001)
Surgery	1.39 (0.05)	1.41 (0.05)	0.76 (0.18)	0.78(0.23)
Center	1.37 (0.11)	1.84 (0.01)	1.29 (0.20)	1.75(0.02)

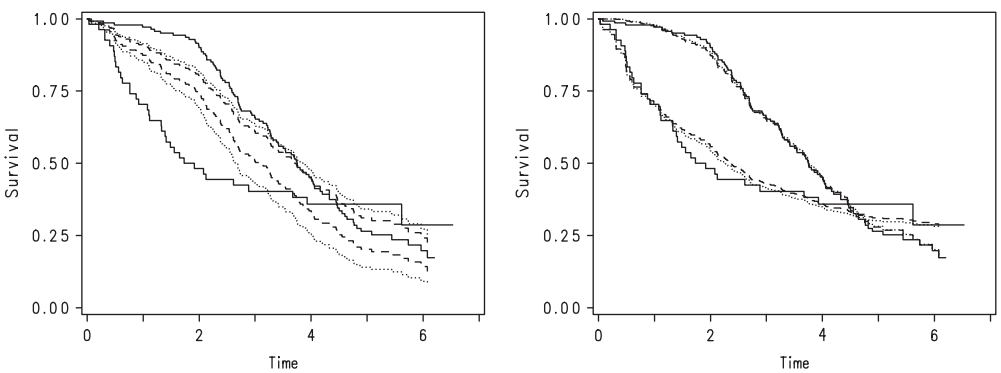

Figure 4. IALT study: left panel: survival functions for centers derived from weighted '·····' and unweighted '– – –' estimation are superimposed to corresponding Kaplan–Meier estimates '——' (left panel); same as left but estimation includes terms for $\log t$ time-dependent effects (right panel).

by neither approach agree well with the Kaplan–Meier estimates, those from WCR getting slightly closer to the Kaplan–Meier estimates for earlier times at the expense of a worse fit for later times. A better local fit in time can only be obtained by introducing time-dependent effects terms.

DOI: 10.1002/sim

This is shown in the right panel of Figure 4 where the fit by either approach is satisfactory. Of course, we cannot expect substantial advantages of WCR of better modeling time dependence. The main goal of this approach lies in a better summary of the average effect.

6. FURTHER REMARKS

The weighted estimation and inference approach to Cox regression which we have presented generalizes the k-sample weighted logrank tests of the Tarone and Ware [35] scheme (i.e. the tests by Breslow [36], Tarone and Ware [35], Prentice [14] and Harrington and Fleming [37]) to a model with multiple, binary and continuous covariates. Our interest has focused, however, on the weighted approach related to Prentice [14], which permits to obtain well interpretable average hazard ratios, independent of proportionality. For average hazard ratios not too large, say up to 4 with a binary predictor, $\exp(\hat{\beta})$ from WCR (approximately) estimates the OC statistic and thus $P(T_1<T_0)$. The latter measure permits a clear decision on which treatment or level of a prognostic factor is preferable, again independent of the proportionality of hazards.

Under proportionality the unweighted and the weighted estimators have the same (approximate) OC interpretation. However, results presented in Section 4 suggest that under non-proportional hazards satisfactory approximation of OC can only be achieved by the weighted estimator. Thus, the weighted approach also provides a possible generalization of OC to a multiple regression model.

Under proportional hazards all weighted approaches will entail some loss of efficiency: we have quantified this effect with the proportional hazards situations A and D in Table II where variances of the weighted approach slightly exceed those from the unweighted approach. From theoretical calculations of efficiency we know that, with increasing censoring, efficiencies will become even more similar. Furthermore, ideal proportional hazards are rarely observed in practice. Therefore, if 'proportional' covariates are analyzed by the weighted approach, the expected loss in efficiency should be quite small. On the other hand, if 'non-proportional' covariates are analyzed by this approach, the intuitive interpretability of effects is maintained and, under converging hazards, the most frequent violation of proportionality, also power is higher than under the unweighted approach, as shown for situation B in Table III.

The weighted approach reduces the complexity of an analysis, in particular, if time-dependencies of continuous covariates [6] or of interactions are observed. Thus, simplification of results can be the motivation for using weighted estimation—independent of the size of samples analyzed. Of course, if available sample sizes are small and/or the numbers of predictors large, then the attractiveness of the weighted approach further increases. In this case, and in particular in its extreme form with modeling under the 'p (# predictors)$\gg n$ (# events)' paradigm, e.g. within penalized regression, advantages of a more robust Cox-type analysis are obvious, as has been pointed out by Gui and Li [40]. In a similar vein, Sasieni [18] confirmed improved power with contaminated data sets and Valsecchi, Silvestri and Sasieni [41] reported improved robustness of weighted estimation. The performance of weighted estimation in the context of the analysis of gene-expression data from microarray studies is currently being investigated by our team.

Approaches for taking account of time-dependence by additional parameters have been an area of intensive research. They usually result in more complex models. Weighted estimation, on the contrary, can simplify analysis and interpretation. In addition, if there is little medical interest in the time-dependence of a covariate effect, a direct comparison of the magnitude of effects from

DOI: 10.1002/sim

'proportional' and 'non-proportional' covariates of interest, a more robust analysis required, or simply, if the robustness of results derived under CR is to be checked, then the weighted estimation approach may be the method of choice.

For CR several extensions have been developed and proven to be useful in practice. These techniques are also available for a weighted analysis. For the IALT data set, e.g. we have presented survival functions based on the weighted approach, along with a corresponding value of explained variation. Similarly, one can obtain Schoenfeld and dfbeta residuals under the weighted approach, include qualitative factors for stratification, model time-dependent covariates and time-dependent effects, and employ the sandwich estimator for taking into account possible dependencies of the observations in a data set. These techniques have been implemented in a SAS macro *WCM* and in an R package *coxphw* for weighted estimation in Cox regression, available at www.muw.ac.at/msi/biometrie/programs.

APPENDIX A: IDENTITY OF AVERAGE HAZARD RATIO DEFINITIONS OC AND $\mathrm{AHR}_{\mathrm{OC}}$

The average hazard ratio definition by Kalbfleisch and Prentice [21]

$$\mathrm{AHR}=\frac{\int(h_1(t)/h(t))f(t)w(t)\,\mathrm{d}t}{\int(h_0(t)/h(t))f(t)w(t)\,\mathrm{d}t}$$

with $h_1(t)=f_1(t)/S_1(t)$, $h(t)=h_0(t)+h_1(t)$, $f(t)=(f_0(t)+f_1(t))/2$ and $w(t)=(S_0(t)f_1(t)+S_1(t)f_0(t))/(f_0(t)+f_1(t))$ simplifies to

$$\mathrm{AHR}_{\mathrm{OC}}=\frac{\int S_0(t)f_1(t)\,\mathrm{d}t}{\int S_1(t)f_0(t)\,\mathrm{d}t}$$

This expression can be rewritten as

$$\mathrm{AHR}_{\mathrm{OC}}=\frac{\int P(T_0>t)f_1(t)\,\mathrm{d}t}{\int P(T_1>t)f_0(t)\,\mathrm{d}t}=\frac{P(T_0>T_1)}{P(T_1>T_0)}=\frac{P(T_1<T_0)}{1-P(T_1<T_0)}=\mathrm{OC}$$

ACKNOWLEDGEMENTS

The authors are grateful to Ariane Dunant and Thierry Le Chevalier, Institut Gustave-Roussy, Villejuif, France, and the IALT Collaborative Group for providing the IALT data set. They also want to thank Robin Henderson, Newcastle, U.K., and Daniela Dunkler, Vienna, Austria, for helpful suggestions on an earlier version of this article, and Terry Smith, Kansas City, U.S.A., for correcting the manuscript. We would also like to thank two anonymous referees for their valuable suggestions.

The project is supported by grant P18553-N13 of the FWF Austrian Science Fund.

REFERENCES

1. Cox DR. Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:187–220.
2. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine* 1994; **13**:1045–1062. DOI: 10.1002/sim.4780131007.
3. Gray RJ. Spline-based tests in survival analysis. *Biometrics* 1994; **50**:640–652. DOI: 10.2307/2532779.

4. Quantin C, Abrahamowicz M, Moreau T, Bartlett G, MacKenzie T, Tazi MA, Lalonde L, Faivre J. Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. *American Journal of Epidemiology* 1999; **150**:1188–1200.
5. Berger U, Schäfer J, Ulm K. Dynamic Cox modelling based on fractional polynomials: time-variations in gastric cancer prognosis. *Statistics in Medicine* 2003; **22**:1163–1180. DOI: 10.1002/sim.1411.
6. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine* 2007; **26**:392–408. DOI: 10.1002/sim.2519.
7. Lehr S, Schemper M. Parsimonious analysis of time-dependent effects in the Cox model. *Statistics in Medicine* 2007; **26**:2686–2698. DOI: 10.1002/sim.2742.
8. Bennett S. Analysis of survival data by the proportional odds model. *Statistics in Medicine* 1983; **2**:273–277. DOI: 10.1002/sim.4780020223.
9. Pettitt AN. Proportional odds models for survival data and estimates using ranks. *Applied Statistics* 1984; **33**:169–175.
10. Murphy SA, Rossini AJ, Van der Vaart AW. Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* 1997; **92**:968–976.
11. Bennett S. Log-logistic regression models for survival data. *Applied Statistics* 1983; **32**:165–171.
12. Marubini E, Valsecchi MG. *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley: New York, 1995.
13. Collett D. *Modelling Survival Data in Medical Research* (2nd edn). Chapman & Hall: London, 2003.
14. Prentice RL. Linear rank tests with right censored data. *Biometrika* 1978; **65**:167–179. DOI: 10.1093/biomet/65.1.167.
15. Wei G, Schaubel DE. Estimating cumulative treatment effects in the presence of nonproportional hazards. *Biometrics* 2008; **64**:724–732. DOI: 10.1111/j.1541-0420.2007.00947.x.
16. Lin DY. Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *Journal of the American Statistical Association* 1991; **86**:725–728.
17. Schemper M. Cox analysis of survival data with non-proportional hazard functions. *The Statistician* 1992; **41**:455–465.
18. Sasieni P. Maximum weighted partial likelihood estimators for the Cox model. *Journal of the American Statistical Association* 1993; **88**:144–152.
19. Sasieni P. Some new estimators for Cox regression. *The Annals of Statistics* 1993; **21**:1721–1759.
20. Xu R, O'Quigley J. Estimating average regression effect under non-proportional hazards. *Biostatistics* 2000; **1**:423–439. DOI: 10.1093/biostatistics/1.4.423.
21. Kalbfleisch JD, Prentice RL. Estimation of the average hazard ratio. *Biometrika* 1981; **68**:105–112. DOI: 10.1093/biomet/68.1.105.
22. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**:515–526. DOI: 10.1093/biomet/81.3.515.
23. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics* 1945; **1**:80–83.
24. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 1947; **18**:50–60.
25. Hanley JA, McNeil BJ. The meaning and use of the area under an ROC curve. *Radiology* 1982; **143**:29–36.
26. Harrell Jr FE. *Regression Modeling Strategies—with Applications to Linear Models*, *Logistic Regression*, *and Survival Analysis*. Springer: New York, 2001.
27. Begun JM, Reid N. Estimating the relative risk with censored data. *Journal of the American Statistical Association* 1983; **78**:337–341.
28. Struthers CA, Kalbfleisch JD. Misspecified proportional hazard models. *Biometrika* 1986; **73**:363–369. DOI: 10.1093/biomet/73.2.363.
29. Wakounig S. *Investigations of the weighted estimation approach for Cox's model. Technical Report 2*, Section of Clinical Biostatistics, Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Vienna, Austria, 2008.
30. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* 1989; **84**:1074–1078.
31. Therneau TM, Grambsch PM. *Modeling Survival Data*. Springer: New York, 2000.
32. Schemper M. A survey of permutation tests for censored survival data. *Communications in Statistics—Theory and Methods* 1984; **13**:1655–1665.
33. Lehmann EL. *Nonparametrics—Statistical Methods Based on Ranks*. Holden Day: San Francisco, 1975.

34. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* 1966; **50**:163–170.
35. Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. *Biometrika* 1977; **64**:156–160. DOI: 10.1093/biomet/64.1.156.
36. Breslow NE. A generalized Kruskal–Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika* 1970; **57**:579–594. DOI: 10.1093/biomet/57.3.579.
37. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika* 1982; **69**:133–143. DOI: 10.1093/biomet/69.3.553.
38. The International Adjuvant Lung Cancer Trial Collaborative Group. Cisplatin-based adjuvant chemotherapy in patients with completely resected non-small-cell lung cancer. *New England Journal of Medicine* 2004; **350**: 351–360. DOI: 10.1056/NEJMoa031644.
39. Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. *Biometrics* 2000; **56**:249–255. DOI: 10.1111/j.0006-341X.2000.00249.x.
40. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 2005; **21**:3001–3008. DOI: 10.1093/bioinformatics/bti422.
41. Valsecchi MG, Silvestri D, Sasieni P. Evaluation of long-term survival: use of diagnostics and robust estimators with Cox's proportional hazards model. *Statistics in Medicine* 1996; **15**:2763–2780. DOI: 10.1002/(SICI)1097-0258(19961230)15:24⟨2763::AID-SIM319⟩3.0.CO;2-O.