Harvard University

Harvard University Biostatistics Working Paper Series

Year 2003 *Paper* 3

The Effects of Misspecifying Cox's Regression Model on Randomized Treatment Group Comparisons

Greg DiRienzo*

*Harvard University, dirienzo@sdac.harvard.edu

Handbook of Statistics, Vol. 23
ISSN: 0169-7161

DOI 10.1016/S0169-7161(03)23022-4

22

The Effects of Misspecifying Cox's Regression Model on Randomized Treatment Group Comparisons

A.G. DiRienzo and S.W. Lagakos

1. Introduction

Hypothesis tests arising from Cox's proportional hazards model (Cox, 1972) are often used to compare randomized treatment groups with respect to the distribution of a failure time outcome. Some of these tests adjust for covariates that may be predictive of outcome, while others, and most notably, the log-rank test, do not. In addition to adjusting for any imbalances that may arise between treatment groups, covariate-adjusted test may enjoy greater efficiency than that of the log-rank test. Tsiats et al. (1985) demonstrated the gain in efficiency of covariate-adjusted tests relative to the log-rank test when the working proportional hazards model is properly specified. Slud (1991) provided asymptotic relative efficiency formulae of the log-rank test to the optimal score test that arises from a properly specified model for covariates when the effect of treatment is multiplicative on the survival time hazard function. Lagakos and Schoenfeld (1984) studied the effects of various types of model misspecification on the power of tests based on Cox's model.

An important consideration in the application of these tests is their validity when the underlying proportional hazards working model is misspecified. Recent work has shown that the impact of model misspecification on the validity of resulting tests hinges on whether the distribution of the potential censoring time either (i) is conditionally independent of treatment group given covariates or conditionally independent of covariates given treatment group, or (ii) depends on both treatment group and covariates. In the first case, resulting test statistics have an asymptotic normal distribution with mean zero under the null hypothesis and that consistent variance estimates are readily obtainable (see Kong and Slud, 1997 and DiRienzo and Lagakos, 2001a). In the second case, the asymptotic mean of the test statistic is not necessarily equal to zero under the null hypothesis when the proportional hazards working model is misspecified. In such cases, the bias of tests can be large, as was demonstrated in DiRienzo and Lagakos (2001a, 2001b).

In this chapter we summarize the properties of hypothesis tests derived from proportional hazards regression models. We introduce notation and define uncorrected statistics in Section 2. In Section 3 we describe conditions necessary for the asymptotic

hs23 v.2003/09/19 Prn:2/10/2003; 13:08 F:hs23022.tex; VTEX/ML p. 2
aid: 23022 pii: S0169-7161(03)23022-4 docsubty: REV

validity of these test statistics, and also discuss efficiency considerations and effects of model misspecification on the power of uncorrected test statistics. We describe a class of corrected test statistics for use when censoring depends on both treatment group and covariates in Section 4, and also examine estimation procedures and the efficiency of such bias-corrected tests. We provide some recommendations for the use of these tests in Section 5, and give MATLAB code for the computation of the various test statistics in Appendix A.

2. Notation and statistics

Let the continuous random variable T denote time from randomization to failure and let C denote a potential censoring time. Assume that we observe $T^* = \min(T, C)$ and the indicator $\delta = \mathbb{1}(T \leqslant C)$ of whether T is observed ($\delta = 1$) or right-censored ($\delta = 0$). Let the binary random variable X denote treatment group and let W denote a $q \times 1$ vector of bounded baseline covariates. Throughout this paper we assume that censoring acts non-informatively, that is, $T \perp C \mid (X, W)$, and also that $X \perp W$, as is the case in most randomized clinical trials. The true conditional hazard functions of T and C given (X, W) are denoted by $\kappa(t \mid X, W)$ and $\kappa_C(t \mid X, W)$, respectively, and are not necessarily of a proportional hazards form. The observed data is assumed to consist of n independent and identically distributed realizations of (T^*, δ, X, Z^*), denoted $(T_i^*, \delta_i, X_i, Z_i^*)$ for $i = 1, \ldots, n$, where Z^* is a $p \times 1$ vector whose components are bounded functions of W.

The null hypothesis of interest is H_0: $X \perp T \mid W$; that is, that the failure time distribution does not depend on treatment group, under which we will denote $\kappa(t \mid X, W)$ by $\kappa(t \mid W)$. Consider tests of H_0 that are based on statistics of the form

$$n^{-1/2}U_n = \sum_{i=1}^{n} \int_0^\infty n^{-1/2} G_n(t)\{X_i - \mathcal{E}_n(t)\}\,\mathrm{d}N_i(t), \tag{1}$$

where

$$\mathcal{E}_n(t) = \sum_{j=1}^{n} Y_j(t)\psi_n(Z_j)X_j \Big/ \sum_{j=1}^{n} Y_j(t)\psi_n(Z_j),$$

$Y_i(t) = \mathbb{1}(T_i^* \geqslant t)$, $N_i(t) = \delta_i \mathbb{1}(T_i^* \leqslant t)$ and $\psi_n(\cdot)$ is a nonrandom bounded function whose form is known but whose parameters can be estimated from the data. The covariates Z_i are some bounded function of Z_i^*, $i = 1, \ldots, n$. The bounded predictable process $G_n(\cdot)$ is also assumed to be nonrandom, converging uniformly in probability to a bounded function $G(\cdot)$. It may be the case that one would want to consider time-dependent covariates, for example an external ancillary covariate process (Kalbfleisch and Prentice, 1980, p. 123). Although results hold when the components of W_i and Z_i are uniformly bounded and predictable functions of time, for ease of notation we only consider fixed covariates.

hs23 v.2003/09/19 Prn:2/10/2003; 13:08 F:hs23022.tex; VTEX/ML p. 3
aid: 23022 pii: S0169-7161(03)23022-4 docsubty: REV

Statistics of the form in (1) arise as the numerator of partial likelihood score tests of $\alpha = 0$ based on working proportional hazards models for $\kappa(t \mid X,\ W)$ that take the form

$$\exp(\alpha X_i)\psi(\beta; Z_i)h(t). \tag{2}$$

The working model (2) is misspecified when it is not equivalent to $\kappa(t \mid X,\ W)$, in which case the parameters α, β, and $h(\cdot)$ have no simple interpretation. The statistic $n^{-1/2}U_n$ should then generally be viewed simply as statistic from which tests of H_0 may be derived.

Popular choices for $G(\cdot)$ and $\psi(\cdot)$ are $G_n(t) = 1$ and $\psi(\beta; Z_i) = \exp(\beta^\top Z_i)$, resulting in $\psi_n(Z_i) = \exp(\hat{\beta}^\top Z_i)$, where $\hat{\beta}$ is the restricted maximum partial likelihood estimator of β obtained by fitting the model with $\alpha = 0$ (Cox, 1972). Here the probability limit of $\psi_n(Z)$ is $\exp(\tilde{\beta}^\top Z)$, where $\tilde{\beta}$ is the probability limit of $\hat{\beta}$ (Lin and Wei, 1989). Note that $\tilde{\beta} = \beta$ when the model (2) is properly specified. Another special case of (1) is the class of weighted log-rank statistics (Cox and Oakes, 1984, p. 124), where $\psi(\beta; Z_i) = \psi_n(Z_i) = 1$, and where the most commonly used choice for $G_n(\cdot)$ is the identity function, yielding the ordinary log-rank test. In general, $\psi(\beta; Z)$ can also depend on t, so long as it is uniformly bounded.

3. Conditions for valid tests

Suppose that either $C \perp X \mid W$ or $C \perp W \mid X$. Then the test statistic $n^{-1/2}U_n$ has an asymptotic normal distribution with mean 0 under H_0, regardless of whether or not the model (2) is misspecified (DiRienzo and Lagakos, 2001a). Furthermore, when either of these conditions hold, consistent estimates of the variance of $n^{-1/2}U_n$ are easily derived, yielding asymptotically valid inference whether or not the relationship between T and (X, W) is properly specified.

The condition $C \perp X \mid W$ is usually satisfied in a randomized clinical trial when the only form of censoring is administrative or end-of-study censoring; that is, when C represents the time from enrollment of a subject into the study until the time the data are analyzed. However, when censoring can arise from premature study discontinuation or loss-to-follow-up, it is well known that this condition may not hold. The condition $C \perp W \mid X$ holds when there is a dependency of censoring on treatment group which does depend on the covariates. To provide some insight into why either of these conditions are necessary for valid inference, note that at baseline (that is, when $t = 0$), the distribution of W is independent of X because of randomization; when either $C \perp X \mid W$ or $C \perp W \mid X$ holds and H_0 is true, it is implied that $X \perp W \mid Y(t) = 1$, $t > 0$, which is necessary for $n^{-1/2}U_n$ to have mean 0 asymptotically. For a proof of these results, see Appendix A of DiRienzo and Lagakos (2001a) or Kong and Slud (1997).

We now provide test statistics for use when either $C \perp X \mid W$ or $C \perp W \mid X$ holds. It follows from Kong and Slud (1997) that under H_0, $n^{-1/2}U_n$ can be expressed as

$$n^{-1/2}U_n = n^{-1/2}\sum_{i=1}^{n} Q_i + o_p(1),$$

where

$$Q_i = \int_0^\infty \{X_i - \mu(t)\}\{dN_i(t) - \rho(t)Y_i(t)\psi(\tilde{\beta}; Z_i)\,dt\},$$

$$\mu(t) = E\{Y(t)\psi(\tilde{\beta}; Z)X\}/E\{Y(t)\psi(\tilde{\beta}; Z)\},$$

$$\rho(t) = E\{Y(t)\kappa(t \mid W)\}/E\{Y(t)\psi(\tilde{\beta}; Z)\}.$$

It is easily verified (cf. Kong and Slud, 1997 or using Lemmas 1 and 2 in DiRienzo and Lagakos, 2001a) that under H_0, Q_i has mean 0 when $C \perp X \mid W$ or $C \perp W \mid X$. This implies that $n^{-1/2}U_n$ is asymptotically normal with mean zero and variance equal to the variance of Q_i. As shown in Kong and Slud (1997), a consistent estimate of the variance of $n^{-1/2}U_n$ is

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{Q}_i - \overline{Q})^2,$$

where

$$\widehat{Q}_i = \int_0^\infty \{X_i - \overline{X}(t)\}\left\{dN_i(t) - \frac{Y_i(t)\psi(\hat{\beta}; Z_i)}{\sum_{j=1}^n Y_j(t)\psi(\hat{\beta}; Z_j)}\,d\overline{N}(t)\right\},$$

$$\overline{X}(t) = \sum_{i=1}^{n} Y_i(t)X_i \Big/ \sum_{i=1}^{n} Y_i(t),$$

$\overline{N}(t) = \sum_{i=1}^n N_i(t)$ and $\overline{Q} = (1/n)\sum_{i=1}^n \widehat{Q}_i$.

Thus, provided that $C \perp X \mid W$ or that $C \perp W \mid X$, the test statistic $U_n/\sqrt{\{\sum_{i=1}^n(\widehat{Q}_i - \overline{Q})^2\}}$ is asymptotically standard normal under H_0 when $C \perp X \mid W$ or $C \perp W \mid X$, regardless of whether the working model (2) is properly specified. The motivation for replacing $\mu(t)$ with $\overline{X}(t)$ above is that $\mu(t) = E\{X \mid Y(t) = 1\}$ under H_0 when $C \perp X \mid W$ or $C \perp W \mid X$. We note that for the special case of the log-rank test, use of the model-based variance estimator of $n^{-1/2}U_n$ also results in a valid asymptotic test, and appears to usually provide nominal finite-sample type I errors (see DiRienzo and Lagakos, 2001a), so that use of a robust variance estimator is not needed.

3.1. *Efficiency considerations*

Lagakos and Schoenfeld (1984) investigated the effects of various types of misspecification of the working model (2) on the power of $n^{-1/2}U_n$. When covariates have a multiplicative effect on the true hazard $\kappa(t \mid X, W)$, but the ratio $\kappa(t \mid X = 1, W)/\kappa(t \mid X = 0, W)$, is non-constant but either greater or less than one for all $t > 0$, i.e., the hazards do not cross, there is often only a small loss in power. One exception to this is when the ratio $\kappa(t \mid X = 1, W)/\kappa(t \mid X = 0, W)$ departs from one only after the majority of failures have occurred; in this case, the loss in power can be great. In contrast, when the ratio $\kappa(t \mid X = 1, W)/\kappa(t \mid X = 0, W)$ crosses one, the loss in power is often substantial.

Suppose that the effect of covariates in the true model $\kappa(t \mid X,\ W)$ is not multiplicative, that is the ratio $\kappa(t \mid X = 1,\ W)/\kappa(t \mid X = 0,\ W)$ is a function of W, but that the interaction is qualitative, in the sense that $\kappa(t \mid X = 1,\ W)/\kappa(t \mid X = 0,\ W)$ is either greater or less than one for all W. In this case, the loss in the power of $n^{-1/2}U_n$ is not in general large unless the discrepancy in the ratio $\kappa(t \mid X = 1,\ W)/\kappa(t \mid X = 0,\ W)$ between levels of W is substantial, especially if larger ratios tend to occur within levels of W that are less prevalent.

More generally, the loss in power of $n^{-1/2}U_n$ can be large when a component of W that has a strong effect on the hazard of T is either omitted or mismodeled in such a way that the direction of its effect is not maintained. Further details on all of these situations can be found in Lagakos and Schoenfeld (1984). Morgan (1986) provides a correction to Lagakos and Schoenfeld's (1984) asymptotic relative efficiency formula of the log-rank test to the score test arising from a properly specified model for covariates. See also Lagakos (1988), who derived asymptotic relative efficiency formulae in the one-sample problem when evaluating the effect of a misspecified form of a time-dependent covariate.

4. Bias correction

When the distribution of the censoring variable depends on both treatment group and covariates, that is when the conditions $C \perp X \mid W$ and $C \perp W \mid X$ both fail to hold, the statistic $n^{-1/2}U_n$ in general has a non-zero asymptotic mean under H_0. One exception is when the model (2) is equal to $\kappa(t \mid X,\ W)$, i.e., the working proportional hazards model is properly specified. DiRienzo and Lagakos (2001a, 2001b) present simulation results which demonstrate that the bias of tests based on $n^{-1/2}U_n$ can be severe when in this setting and the working proportional hazards model is misspecified.

In an attempt to correct for this bias, DiRienzo and Lagakos (2001b) present a class of tests that are asymptotically standard normal under H_0 regardless of the joint distribution between C and $(X,\ W)$, provided that either the conditional distribution of T given $(X,\ W)$ or the conditional distribution of C given $(X,\ W)$ is properly modeled. Consequently, these tests are more robust than those arising from $n^{-1/2}U_n$ when the working model is misspecified, and do not appear to lose much efficiency when the working model is correctly specified and bias correction is unnecessary.

Consider the generalization of (1) given by

$$n^{-1/2}U_n^* = \sum_{i=1}^{n} \int_0^\infty n^{-1/2} G_n(t)\varphi(t; X_i, W_i)\big\{X_i - \mathcal{E}_n^*(t)\big\}\,\mathrm{d}N_i(t), \tag{3}$$

where

$$\mathcal{E}_n^*(t) = \sum_{j=1}^{n} Y_j^*(t)\psi_n(Z_j)X_j \Big/ \sum_{j=1}^{n} Y_j^*(t)\psi_n(Z_j),$$

$$Y_i^*(t) = Y_i(t)\varphi(t; X_i, W_i),$$

$$\varphi(t; X_i, W_i) = \min\big\{ \operatorname{pr}(C \geqslant t \mid X_i = 0,\ W_i), \operatorname{pr}(C \geqslant t \mid X_i = 1,\ W_i)\big\} / \operatorname{pr}(C \geqslant t \mid X_i,\ W_i), \tag{4}$$

for $i = 1, \ldots, n$. Unlike the binary indicator variable Y normally used in Cox's model, $Y_i^*(t)$ can assume any value in the unit interval. Also, note that $\varphi(t; X_i,\ W_i)$ is only defined when $Z_i = W_i$, $i = 1, \ldots, n$.

At each point in study time when a survival event occurs, this correction strives to remove any imbalances between treatment groups in the distribution of covariates that are caused solely by censoring. Mechanically, at study time t, the correction down-weights, $Y_i^*(t) < 1$, those subjects in the risk set whose risk of censoring is higher in their opposite treatment group; those subjects whose risk of censoring is lower in their opposite treatment group are unweighted, $Y_i^*(t) = Y_i(t) = 1$. To see this analytically, note that under H_0, the conditional expectation of $Y^*(t)$ given (X, W) is

$$\begin{aligned}
&\varphi(t; X,\ W) \operatorname{pr}\{Y(t) = 1 \mid X,\ W\} \\
&\quad = \varphi(t; X,\ W) \operatorname{pr}(C \geqslant t \mid X,\ W) \operatorname{pr}(T \geqslant t \mid W) \\
&\quad = \min\big\{\operatorname{pr}(C \geqslant t \mid X = 0,\ W), \operatorname{pr}(C \geqslant t \mid X = 1,\ W)\big\} \operatorname{pr}(T \geqslant t \mid W),
\end{aligned}$$

which is independent of X. The probability limit of $\mathcal{E}_n^*(t)$ under H_0 is thus

$$\frac{E\{Y^*(t)\psi(Z)X\}}{E\{Y^*(t)\psi(Z)\}} = \frac{E[X\psi(Z)E\{Y^*(t) \mid W\}]}{E[\psi(Z)E\{Y^*(t) \mid W\}]} = \pi,$$

where $\pi = E(X)$.

As shown in DiRienzo and Lagakos (2001b), $n^{-1/2}U_n^*$ can be expressed under H_0 as

$$n^{-1/2}U_n^* = n^{-1/2} \sum_{i=1}^{n} A_i + \mathrm{o}_p(1),$$

where

$$\begin{aligned}
A_i = &\int_0^\infty G(t)\varphi(t; X_i, W_i)(X_i - \pi) \\
&\times \left\{ \mathrm{d}N_i(t) - Y_i(t)\psi(Z_i) \frac{E\{Y^*(t)\kappa(t \mid W)\}}{E\{Y^*(t)\psi(Z)\}} \,\mathrm{d}t \right\},
\end{aligned}$$

and the A_i are independent and identically distributed with mean zero.

A consistent estimator of the variance of $n^{-1/2}U_n^*$ is

$$V_n = \frac{1}{n} \sum_{i=1}^{n} \big(A_i^{(n)} - \overline{A}^{(n)}\big)^2, \tag{5}$$

where

$$A_i^{(n)} = \int_0^\infty G_n(t)\varphi(t; X_i, W_i)\big(X_i - \overline{X}\big) \times \left\{ \mathrm{d}N_i(t) - \frac{Y_i(t)\psi_n(Z_i)}{\sum_{j=1}^n Y_j^*(t)\psi_n(Z_j)} \sum_{j=1}^n \varphi(t; X_j, W_j)\,\mathrm{d}N_j(t) \right\},$$

$\overline{X}$ is the mean of $\{X_1, \ldots, X_n\}$ and $\overline{A}^{(n)}$ is the mean of $\{A_1^{(n)}, \ldots, A_n^{(n)}\}$. Hence, regardless of the joint distribution between C and $(X,\ W)$, $n^{-1/2}U_n^*/\sqrt{V_n}$ asymptotically has the standard normal distribution under H_0 whether or not the working model is properly specified. It follows that if the working model (2) is properly specified, $n^{-1/2}U_n^*/\sqrt{V_n}$ is asymptotically standard normal under H_0 regardless of whether $\mathrm{pr}(C \geqslant t \mid X,\ W)$ is properly specified and of the dependency between C and (X, W).

In practice, $\varphi(\cdot)$ will often be unknown. Let $\hat{\varphi}(t; X_i,\ W_i)$ denote an estimator of $\varphi(t; X_i, W_i)$. One would then calculate

$$\widehat{U}_n^* = \sum_{i=1}^n \int_0^\infty G_n(t)\hat{\varphi}(t; X_i, W_i)\big\{X_i - \widehat{\mathcal{E}}_n^*(t)\big\}\,\mathrm{d}N_i(t)$$

instead of (3) and

$$\widehat{A}_i^{(n)} = \int_0^\infty G_n(t)\hat{\varphi}(t; X_i, W_i)\big(X_i - \overline{X}\big) \times \left\{ \mathrm{d}N_i(t) - \frac{Y_i(t)\psi_n(Z_i)}{\sum_{j=1}^n \widehat{Y}_j^*(t)\psi_n(Z_j)} \sum_{j=1}^n \hat{\varphi}(t; X_j, W_j)\,\mathrm{d}N_j(t) \right\}$$

instead of $A_i^{(n)}$ in (5), where $\widehat{Y}_i^*(t) = Y_i(t)\hat{\varphi}(t; X_i,\ W_i)$, and $\widehat{\mathcal{E}}_n^*(\cdot)$ is obtained by substituting $\widehat{Y}_i^*(\cdot)$ for $Y_i^*(\cdot)$ in $\mathcal{E}_n^*(\cdot)$, $i = 1, \ldots, n$. Denote this variance estimate by $\widehat{V}_n$.

Some methods for estimating $\varphi(t; X, W)$ are given in DiRienzo and Lagakos (2001b). These include the nonparametric regression methods of McKeague and Utikal (1990) as well as Cox's (1972) proportional hazards regression models. If the covariates are discrete with relatively few levels, then a stratified, left-continuous Kaplan–Meier estimator (Kaplan and Meier, 1958) of censoring can be calculated for each treatment group within each level of the covariate space.

For example, an estimate for $\varphi(t; X, W)$ can be obtained via the stratified proportional hazards model for $\kappa_C(t \mid X,\ W)$,

$$\lambda^{(X)}(t)\exp\big(\gamma^{(X)}Z^C\big),$$

where Z_i^C is some bounded function of Z_i^*, $i = 1, \ldots, n$. The maximum partial-likelihood estimator, $\hat{\gamma}_t^{(X)}$, and the Breslow (1972, 1974) estimator of the baseline cumulative hazard function of censoring, $\widehat{\Lambda}^{(X)}(t)$, may then be calculated within each treatment group at each censoring time, using data accumulated before that time, and the continuous estimator

$$\widehat{\mathrm{pr}}\big(C_i \geqslant t \mid X_i,\ Z_i^C\big) = \exp\big\{-\widehat{\Lambda}^{(X_i)}(t)\exp\big(\hat{\gamma}_t^{(X_i)}Z_i^C\big)\big\}$$

hs23 v.2003/09/19 Prn:2/10/2003; 13:08 F:hs23022.tex; VTEX/ML p. 8
aid: 23022 pii: S0169-7161(03)23022-4 docsubty: REV

obtained by linear interpolation between censoring times of $\widehat{\Lambda}_i^{(X_i)}(t)$. Here estimation was stratified on X, but stratification may additionally be based on any covariate that might possibly have a strong interaction with treatment.

When $\varphi(t; X, W)$ is estimated using a semiparametric or nonparametric model, $\hat{\varphi}(t; X_i, W_i)$ contains estimates of an infinite dimensional parameter, for which case a consistent estimate of the variance of $n^{-1/2}\widehat{U}_n^*$ would not necessarily be given by $\widehat{V}_n$. However, given the choice for an estimate of $\varphi(t; X, W)$, if it can be shown that $\widehat{U}_n^*$ is asymptotically linear, then the nonparametric bootstrap estimate of variance of $\widehat{U}_n^*$ will be consistent (Gill, 1989). DiRienzo and Lagakos (2001b) have shown via simulation that when using a semiparametric proportional hazards model to calculate $\hat{\varphi}(t; X_i, W_i)$, the variance estimate $\widehat{V}_n$ appears to be adequate.

For any given data set, there is no guarantee that it will be possible to specify and estimate $\varphi(\cdot)$ well enough to make the correction for a misspecified model for T reliable. It is thus of utmost importance to check and validate the fit of both the model for censoring and survival. Some well known techniques for checking the appropriateness of proportional hazards regression models are given in Lin et al. (1993) and Klein and Moeschberger (1997).

A related consideration in the use of bias-adjusted tests are the relative efficiencies. When the working proportional hazards model (2) is properly specified, i.e., equal to $\kappa(t \mid X, W)$, then the uncorrected, fully model-based test of H_0 is asymptotically valid regardless of the dependency between C and (X, W). In this situation, it is of interest to examine the relative efficiency of the corrected test to that of the uncorrected test and determine if there are situations for which unnecessary use of the corrected test could lead to loss in power.

DiRienzo and Lagakos (2001b) provide formulae for the asymptotic mean and variance of $n^{-1/2}\widehat{U}_n$ and $n^{-1/2}\widehat{U}_n^*$ under the contiguous alternative H_n: $\alpha = c/\sqrt{n}$, for some constant c, when the true hazard for T is given by

$$\kappa(t \mid X_i, W_i) = \exp(\alpha X_i)\psi(\beta, W_i)h(t).$$

That is, the working proportional hazards model is properly specified and calculation of a corrected test is unnecessary since the uncorrected, fully model-based test of H_0 is asymptotically valid. In their accompanying simulations, the empirical relative efficiency of the corrected test to that of the uncorrected test appears to almost always be close to one.

Other choices for the functional form of $\varphi(\cdot)$ may be of interest; one example is $\varphi(t; X, W) = 1/\text{pr}(C \geqslant t \mid X, W)$. However, using simulations, DiRienzo and Lagakos (2001b) have found that this choice for $\varphi(\cdot)$ can be much less efficient than the choice (4). They present an efficacy formula for the corrected test; this may be used to compare the efficiencies of tests using different choices for $\varphi(\cdot)$.

5. Discussion

Given the wide use of statistical tests based on Cox's regression model, especially in medical applications, and considering the importance of decisions that are reached from

these analyses, an understanding of their robustness to misspecification of the model is important. Misspecification can occur in many forms, including omitted or mismodelled covariates, the omission of treatment by covariate interactions, or a violation of the underlying proportionality assumption. While goodness-of-fit methods can be applied to check model fit (cf. Klein and Moeschberger, 1997), their failure to signal misspecification is no assurance that this is the case and, furthermore, their subjective and post-hoc nature can be problematic when a new treatment is being assessed, e.g., in clinical trials the standard practice is to precisely prespecify how treatment comparisons will be made. This chapter has argued that a fundamental question in assessing such robustness is whether treatment group and the censoring variable are conditionally independent given the underlying covariates, or whether the underlying covariates associated with survival are conditionally independent of the censoring variable, given treatment group. When either of these conditions apply, then statistical tests arising from fitting a proportional hazards model, including the popular log-rank test, maintain their validity under misspecification of the model-relating treatment and these covariates to the hazard function for survival. That is, when either condition holds, the resulting test statistic, when standardized by a robust variance estimator, has a distribution under the null hypothesis of no treatment effect that is asymptotically standard normal, regardless of whether or not the model is correctly specified. For the special case of the log-rank test, use of the model-based variance estimator to standardize the score statistic arising from the assumed model also leads to the desired asymptotic behavior under the null hypothesis. Thus, establishment of either of these conditions ensures that the size, or Type I error, associated with such tests is not distorted as a result of model misspecification. Moreover, one or both of the conditions can in practice often either be checked empirically or concluded to hold based on the analyst's knowledge of the circumstances that lead to censored observations.

When neither condition holds, that is, when either treatment or the underlying covariate is not conditionally independent of time to censoring, then tests based on fitting a proportional hazards model can be asymptotically biased under the null hypothesis. Since in practice the significance levels used to evaluate these tests invariably resort to their presumed asymptotic normality, the size of such tests can be seriously biased when the working proportional hazards model is misspecified. To avoid or minimize such biases, a class of bias-corrected tests can readily be adapted. These tests require knowledge or estimation of $\phi(t; X, W)$, a function of the conditional distribution of censoring. Based on asymptotic considerations and simulations, the corrected test works well in a variety of settings, even when the estimated form of $\phi(t; X, W)$ is only approximately correct. That is, misspecification of the function $\phi(t; X, W)$ appears to be far less critical for the bias-corrected test than does the misspecification of the underlying hazard model for the uncorrected test. Furthermore, use of a bias-corrected test when one is unnecessary – that is, when the working proportional hazards model happens to be correctly specified – does not appear to result in much loss in efficiency. Thus, when there is any suspicion that the key conditions for robustness may be violated, use of the bias-adjusted tests instead of or as a complement to standard methods is advised. To facilitate the computation for the adjusted tests, Appendix A gives MATLAB code for these and uncorrected tests.

hs23 v.2003/09/19 Prn:2/10/2003; 13:08 F:hs23022.tex; VTEX/ML p. 10
aid: 23022 pii: S0169-7161(03)23022-4 docsubty: REV

Acknowledgement

This work was supported in part from grant AI24643 from the US National Institutes of Health.

Appendix A: MATLAB code for computing statistical tests

We provide below MATLAB code for calculating the uncorrected and corrected score tests presented in this paper. The version of MATLAB used is 5.3.1 (R11.1) along with the Statistics (Version 2.2, R11) and Optimization (Version 2.0, R11) toolboxes.

The uncorrected test is calculated using the model-based variance estimator, which is consistent when the working proportional hazards model for $\kappa(t \mid X,\ W)$ is properly specified or when the log-rank test is used as the uncorrected test. The corrected test is calculated with $G_n(t) = 1$ and using a stratified (by treatment group) proportional hazards model for the conditional distribution of C given (X, W) with $\psi_n(Z) = \exp(\hat{\beta}^\top Z)$, where $\hat{\beta}$ the restricted maximum partial likelihood estimate of β under H_0. We note, however, that the code can be modified to accommodate other choices for these functions as well as for more covariates that are used below to illustrate the methods.

The observed data consists of the five $n \times 1$ column vectors `T0`, `d`, `x`, `Z1`, `Z2`, where `T0` corresponds to $\{T_i^*\}$, `d` to $\{\delta_i\}$, `x` to $\{X_i\}$, `Z1` is the first component of $\{Z_i^*\}$, say $\{Z_{1i}\}$ and `Z2` the second, say $\{Z_{2i}\}$, $i = 1, \ldots, n$. Suppose that one wanted to adjust for the covariates $I(Z_1 < 0)$, Z_2^2 in the model for T, and calculate a corrected test using a proportional hazards model for C that was conditional on $X, |Z_1|^{-1/2}, Z_2$. Then the MATLAB call would be

$$[\texttt{un1}, \texttt{cor1}] = \texttt{SC}(\texttt{T0}, \texttt{d}, \texttt{x}, [(\texttt{Z1} < \texttt{0}), (\texttt{Z2}.\hat{}(\texttt{2}))],$$

$$[((\texttt{abs}(\texttt{Z1})).\hat{}(-.5)), \texttt{Z2}]);$$

where the output 1×2 row vector `un1` consists of the uncorrected score statistic and score test, similarly, `cor1` consists of the corrected score statistic and score test.

The code for the function `SC.m` and the two functions it calls, `rPLgh.m` and `BRES.m` is given by:

```
function  [un, cor] = SC(TT, dd, xx, Z1, Z2)
% computes uncorrected *un* and corrected *cor* score statistics
   and tests
% *TT* is the column vector of N possibly right-censored event
   times
%      there are assumed to be no TIES in *TT*
% *dd* is the column vector of N indicators I(T<=C)
% *xx* is the column vector of N treatment group indicators
% *Z1* is the N x p matrix of covariates for *T*
% *Z2* is the N x p matrix of covariates for *C*
%-------------------------------------------------------------------
```

hs23 v.2003/09/19 Prn:2/10/2003; 13:08 F:hs23022.tex; VTEX/ML p. 11
aid: 23022 pii: S0169-7161(03)23022-4 docsubty: REV


```
% first get the restricted MPLE for T
%-----------------------------------------------------------------
global T d x Z;
T=TT; d=dd; x=xx; Z=Z1;
  N = length(T);
  p = size(Z,2);
th = zeros(1,p);
options = optimset('GradObj','on','Display','off');
  rmple = fsolve('rPLgh',th,options);
clear global
%-----------------------------------------------------------------
% calculate the MPLE and Breslow estimate of the baseline
% cumulative hazard of censoring within each treatment group
%-----------------------------------------------------------------
global T d Z;
T=TT(xx==0); d=1-dd(xx==0); Z=Z2(xx==0,:);
  N = length(T);
  p = size(Z,2);
th = zeros(1,p);
options = optimset('GradObj','on','Display','off');
mple0 = fsolve('rPLgh',th,options);
[L0w,c0] = BRES(T,d,Z,mple0);
clear global
%-----------------------------------------------------------------
global T d Z;
T=TT(xx==1); d=1-dd(xx==1); Z=Z2(xx==1,:);
  N = length(T);
  p = size(Z,2);
th = zeros(1,p);
options = optimset('GradObj','on','Display','off');
mple1 = fsolve('rPLgh',th,options);
[L1w, c1] = BRES(T,d,Z,mple1);
clear global
%-----------------------------------------------------------------
  T = TT; d = dd; x = xx; Z = Z1;
  p = size(Z,2);
  I = zeros((p+1),(p+1));
  U = 0; Ur = 0;
  Ts = T.*d;
  N = length(T);
  K = sum(d);
eBz = exp((Z*(rmple'))');
Xb  = mean(x);
  wr = zeros(N,1); Wr = zeros(N,1);
temp0  = 0; dN = 0; YseM = zeros(N,1);
%-----------------------------------------------------------------
% calculate the baseline cumulative hazards at each observed
% failure time for each treatment group by linear interpolation
%-----------------------------------------------------------------
```

hs23 v.2003/09/19 Prn:2/10/2003; 13:08 F:hs23022.tex; VTEX/ML p. 12
aid: 23022 pii: S0169-7161(03)23022-4 docsubty: REV


```
L0 = interp1(c0,L0w, min(max(c0),T));
L1 = interp1(c1,L1w, min(max(c1),T));
%-----------------------------------------------------------------
% test statistic
%-----------------------------------------------------------------
for mm=1:N
  if (Ts(mm)>0)
     Y = (T>=Ts(mm));
     Y0 = Y.*(1-x);
     Y1 = Y.*x;
%-----------------------------------------------------------------
%treatment-specific Survival functions of censoring
%-----------------------------------------------------------------
    F0 = exp(-L0(mm)*exp(Z2*mple0'));
    F1 = exp(-L1(mm)*exp(Z2*mple1'));
    F0 = F0 + (F0==0).*eps;
    F1 = F1 + (F1==0).*eps;
phi0r = ((min([F1';F0'])./F0')';
phi1r = ((min([F1';F0'])./F1')';
%-----------------------------------------------------------------
% uncorrected test
%-----------------------------------------------------------------
  meBz = (eBz'.*Y)*ones(1,p+1);
    s0 = (eBz*Y)/N;
    s1 = (sum(meBz.*[x,Z]))/N;
    s2 = ([x,Z]'*(meBz.*[x,Z]))/N;
    vz = ((s2/s0) - ((s1/s0)'*(s1/s0)));
    I = I + vz/N;
    sc = ([x(mm),Z(mm,:)] - (s1/s0));
    U = U + sc(1);
%-----------------------------------------------------------------
% corrected test
%-----------------------------------------------------------------
    Y0n=Y0.*phi0r;
    Y1n=Y1.*phi1r;
    Ys = Y0n + Y1n;
    YseM = [YseM, (eBz'.*Ys)];
   meBz = (eBz'.*Ys)*ones(1,p+1);
      s0 = (eBz*Ys)/N;
  temp0 = [temp0, s0];
      s1 = (sum(meBz.*[x,Z]))/N;
      s2 = ([x,Z]'*(meBz.*[x,Z]))/N;
      E = s1/s0;
      Ur = Ur + ( Ys(mm)*(x(mm)-E(1)) );
      dN = [dN, Ys(mm)];
 wr(mm) = Ys(mm)*( x(mm) - Xb );
 end
end
 temp0(1)=[];
```

hs23 v.2003/09/19 Prn:2/10/2003; 13:08 F:hs23022.tex; VTEX/ML p. 13
aid: 23022 pii: S0169-7161(03)23022-4 docsubty: REV


```
     dN(1)=[];
YseM(:,1)=[];
I = N*I;
%-----------------------------------------------------------------
%calculate sample version of the iid terms (A)
%-----------------------------------------------------------------
resr=sum((((x-Xb)*ones(1,K)).*(((YseM)./(ones(N,1)*temp0)).*
(ones(N,1)*(dN/N))))';
  Wr = wr - resr';
%-----------------------------------------------------------------
%model-based variance estimate of uncorrected test
%-----------------------------------------------------------------
aa=I((2:p+1),(2:p+1));
iiI=inv(aa);
% iiI=aa\eye(size(aa)); may be more efficient
V = I(1,1)-(I(1,(2:p+1))*iiI*I((2:p+1),1));
%-----------------------------------------------------------------
%variance estimate of corrected test
%-----------------------------------------------------------------
Rrm =  sum( (Wr-mean(Wr)).^(2) );
  un = [U,  U/sqrt(V)];
cor = [Ur, Ur/sqrt(Rrm)];
%-----------------------------------------------------------------
% NOTE: to calculate log-rank, set rmple=zeros(1,p) and
% V = I(1,1);
function [dL, ddL] = rPLgh(th)
% computes the gradient and Hessian of Cox's partial likelihood
   at *th*
% *th* is the (p+1) row vector of coefficients
% *T* is the column vector of N possibly right-censored event
   times
% *d* is the column vector of N  indicators I(T<=C)
% *Z* is the N-by-p matrix of baseline covariates
global T d Z;
  N = length(T);
  p = size(Z,2);
  I = zeros(p,p);
  U = zeros(1,p);
%-----------------------------------------------------------------
% compute S^0(th,t), S^1(th,t) and S^2(th,t) at each event time
Bz = Z*(th');
eBz = exp(Bz');
Ts = T.*d;
for n=1:N
  if (Ts(n)>0)
     Y = (T>=Ts(n));
     meBz = (eBz'.*Y)*ones(1,p);
        s0 = (eBz*Y)/N;
        s1 = (sum(meBz.*Z))/N;
```

hs23 v.2003/09/19 Prn:2/10/2003; 13:08 F:hs23022.tex; VTEX/ML p. 14
aid: 23022 pii: S0169-7161(03)23022-4 docsubty: REV


```
        s2 = (Z'*(meBz.*Z))/N;
        vz = (s2/s0) - ((s1/s0)'*(s1/s0));
        sc = Z(n,:) - (s1/s0);
    U = U + sc;
    I = I + vz;
  end
end
dL = U';
ddL = -I;
function [LL, tt] = BRES(T,d,z,b)
% computes Breslow's estimate of baseline cumulative baseline
   hazard fn
% *T* is the column vector of N possibly right-censored event
   times
% *d* is the column vector of N indicators I(T<=C)
% assumes no ties in the data
% *z* is the N x p matrix of covariates
% *b*  is the 1 x p vector of regression coefficients
Ts=T.*d;
Ts=Ts(Ts>0);
Ts=sort(Ts);
n=length(Ts);
L=1:n;
eb = exp(z*b');
for mm=1:n
 L(mm) = 1/sum( (T>=Ts(mm)).*eb );
end
tt=[0,Ts'];
LL=[0,cumsum(L)];
```

References

Breslow, N.E. (1972). Discussion of the paper by D.R. Cox. *J. Roy. Statist. Soc. B* **34**, 216–217.

Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B* **34**, 187–220.

Cox, D.R., Oakes, D.O. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

DiRienzo, A.G., Lagakos, S.W. (2001a). Effects of model misspecification on tests of no randomized treatment effect arising from Cox's proportional hazards model. *J. Roy. Statist. Soc. B* **63**, 745–757.

DiRienzo, A.G., Lagakos, S.W. (2001b). Bias correction for score tests arising from misspecified proportional hazards regression models. *Biometrika* **88**, 421–434.

Gill, R.D. (1989). Non and semi-parametric maximum likelihood estimators and the von mises method (Part 1). *Scand. J. Statist.* **16**, 97–128.

Kalbfleisch, J.D., Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

Klein, J.P., Moeschberger, M.L. (1997). *Survival Analysis – Techniques for Censored and Truncated Data*. Springer, New York.

Kong, F.H., Slud, E. (1997). Robust covariate-adjusted log-rank tests. *Biometrika* **84**, 847–862.

hs23 v.2003/09/19 Prn:2/10/2003; 13:08 F:hs23022.tex; VTEX/ML p. 15
aid: 23022 pii: S0169-7161(03)23022-4 docsubty: REV

Lagakos, S.W. (1988). The loss in efficiency from misspecifying covariates in proportional hazards regression models. *Biometrika* **75**, 156–160.

Lagakos, S.W., Schoenfeld, D.A. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* **40**, 1037–1048.

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

Lin, D.Y., Wei, L.J., Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

McKeague, I.W., Utikal, K.J. (1990). Inference for a nonlinear counting process regression model. *Ann. Statist.* **18**, 1172–1187.

Morgan, T. (1986). Omitting covariates from the proportional hazards model. *Biometrics* **42**, 993–995.

Slud, E. (1991). Relative efficiency of the log rank test within a multiplicative intensity model. *Biometrika* **78**, 621–630.

Tsiats, A.A., Rosner, G.L., Tritchler, D.L. (1985). Group sequential tests with censored survival data adjusting for covariates. *Biometrika* **72**, 365–373.