COMPETING RISK AND THE COX PROPORTIONAL HAZARD MODEL

ROGER M. COOKE AND OSWALD MORALES-NAPOLES

Abstract. We propose a heuristic for evaluating model adequacy for the Cox proportional hazard model by comparing the population cumulative hazard with the baseline cumulative hazard. We illustrate how recent results from the theory of competing risk can contribute to analysis of data with the Cox proportional hazard model. A classical theorem on independent competing risks allows us to assess model adequacy under the hypothesis of random right censoring, and a recent result on mixtures of exponentials predicts the patterns of the conditional subsurvival functions of random right censored data if the proportional hazard model holds.

1. Introduction

Recent results in the theory of competing risk involve establishing identifiability of the marginal or competing life variables under a variety of assumptions regarding the censoring mechanism. Each mechanism is associated with a distinctive "footprint" in the subsurvival functions, and these footprints in turn form the basis of statistical tests in testing model adequacy. To date, most applications have been in reliability [8, 13, 9] and biostatistics ([1]). This article shows how these techniques can contribute to the field of proportional hazard modelling ([10], for a recent overview see [19]). The exposition is largely informal. The Cox proportional hazard model is briefly reviewed, with attention to the issue of model adequacy. We propose a simple overall test of adequacy that does not use partial likelihood. Recent results in the theory of dependent competing risk are reviewed in section 4, and we show how these can supplement the diagnostic tools in proportional hazard modelling. Section 5 illustrates these ideas on a lung cancer data set [18]. The final section draws conclusions[1].

2. Proportional Hazard model

To simplify the presentation, we consider the case of time-invariant covariates X, Y, Z without censoring and without ties. We consider data to be generated by the following hazard rate

$$h(X, Y, Z) = \Lambda_0(t)e^{XA+YB+ZC} \tag{2.1}$$

where Λ_0 is the baseline hazard. The covariates (X, Y, Z) are considered as random variables. The coefficients (A, B, C) and the baseline hazard Λ_0 will be estimated

1991 *Mathematics Subject Classification.* 62N01, 62N05,62N99, 62P10.

Key words and phrases. Competing risk, Relative risk, Cox proportional hazard, Censoring, Model adequacy.

[1]The authors gratefully acknowledge many helpful comments from Bo Lindqvist

from life data. If this hazard rate holds, then for an individual with covariate values (x,y,z) the survivor function is

$$e^{-h(x,y,z)}. \tag{2.2}$$

Suppose we observe times of death $t_1, ...t_n$ such that $t_i < t_j$ for $i < j$. Let the covariates for the individual dying at time t_i be denoted (x_i, y_i, z_i). The coefficients A, B, C are estimated by maximizing the *partial likelihood*

$$\prod_{i=1}^{N} \frac{e^{x_i A + y_i B + z_i C}}{\sum_{j \geq i}^{n} e^{x_j A + y_j B + z_j C}} \tag{2.3}$$

Note that the times of death t_i do not appear in (2.3). The intuitive explanation is as follows. *Given* that the first death in the population occurs at time t_1, the probability that it happens to individual 1 is $\frac{e^{x_1 A + y_1 B + z_1 C}}{\sum_{j \geq 1}^{n} e^{x_j A + y_j B + z_j C}}$. After individual 1 is removed from the population, the same reasoning applies to the surviving population; *given* that the second time of death t_2, the probability that it happens to individual 2 is $\frac{e^{x_2 A + y_2 B + z_2 C}}{\sum_{j \geq 2}^{n} e^{x_j A + y_j B + z_j C}}$, and so on. Kalbfleisch and Prentice ([16]) note that for constant covariates, (2.3) is the likelihood for the *ordering* of times of death. The baseline hazard can be estimated from the data as described in ([16]p.114).

3. Model adequacy

Testing model adequacy for the Cox model is not straightforward[2]. In many important studies, model adequacy is not examined, and only individual coefficients for the covariate of interest are reported, with Wald confidence bounds (eg [11, 21]). The coefficients are used to compute relative risk, and form the basis of (dis)utility calculations for different risk mitigation measures.

With (x, y, z) fixed and T random, *and* with constant baseline hazard scaled to one, the survivor function (2.2), considered as a function of the random variable T is uniformly distributed on [0 , 1], that is

$$T \sim -ln(U)/h. \tag{3.1}$$

where U is uniform on $[0, 1]$. As this holds for each individual in the population $i = 1...N$. If we order the values

$$e^{-t_i e^{x_i A + y_i B + z_i C}}; i = 1...N \tag{3.2}$$

[2]This is a sampling of statements found in the literature regarding model evaluation: "it is not apparent what kinds of departures one would expect to see in the residuals if the model is incorrect, or even to what extent agreement with the anticipated line should be expected" ([16], p128). "For most purposes, you can ignore the Cox-Snell and martingale residuals. While Cox-Snell residuals were useful for assessing the fit of the parametric models in Chap ter 4, they are not very informative for the Cox models estimated by partial likelihood" ([2], p 173). "Unfortunately, this distribution theory [of the Cox Snell residuals as exponentially distributed] has not proven to be as useful for model evaluation as the theory derived from the counting process approach". ([15] p. 202), "there is not a single, simple, easy to calculate, useful, easy to interpret measure [of model performance] for a proportional hazards model." ([15] p. 229). "the martingale residuals can not play all the roles that linear model residuals do; in particular the overall distribution of the residuals does not aid in the global assessment of fit." ([22] p 81).

and plot them against their number, the points should lie along the diagonal if the proportional hazard model is true with coefficients A, B, C and constant baseline hazard[3].

This would provide an easy heuristic check of model adequacy if the baseline hazard were indeed known to be constant and scaled to one. However, if the baseline hazard is also estimated from the data, then this simple test does not apply. Thus it may well arise that data generated with a constant baseline hazard appears to acquire a time dependent baseline hazard as a result of missing covariates. Letting $\hat{\diamond}$ denote values estimated from the data, we may well find that the values

$$e^{-\widehat{\Lambda}_0(t_i)e^{x_i\hat{A}+y_i\hat{B}+z_i\hat{C}}}; i = 1...N \tag{3.3}$$

plot as uniform, while the estimates do not equal the values which generated the data. In particular, this may arise in the case of missing covariates. We identify some covariates but many others may not be represented in our model. For example, in considering the influence of airborne fine particulate matter on non-accidental mortality [11, 21], covariates like smoking, sex, age, socio-economic status, air quality, and weather are studied. However time to death is obviously influenced by myriad other factors like occupation, genetic disposition, stress, disease prevalence, medical care, diet, alcohol consumption, home environment (eg radon), travel patterns, etc.etc.

The following type of simple numerical experiment, which the reader may verify for him/herself will illustrate the problems with model adequacy[4].

(1) Choose coefficients (A, B, C), choose a constant baseline hazard scaled to one, and choose a distribution for (X, Y, Z);
(2) Sample independently 100 values of (X, Y, Z) and 100 values from the uniform distribution on $[0, 1]$; compute failure times using (3.1);
(3) Estimate the coefficients by maximizing (2.3), and estimate the baseline hazard.

This procedure does not require that the distributions of the covariates be centered at their means; indeed, centering is not standard procedure in applications. However, the uniform distribution on [-1,1] used here is centered.

Let the model (2.1) be termed h_{XYZ}. To study the effects of model incompleteness estimate the coefficient A with a model h_{XY} using only covariates X and Y, and with a model h_X using only covariate X. For each of the models h_{XYZ}, h_{XY} and h_X, we repeat the above procedure 100 times with the same values for (A, B, C), with (X, Y, Z) sampled independently from the (centered) uniform distribution on $[-1, 1]$. Figure 1 plots the ordered estimates of coefficient A.

Evidently, the models h_{XY} and h_X tend to underestimate the coefficient A. A theoretical explanation of this underestimation is given in [3, 17]. The tendency to underestimate becomes more pronounced in Figure (2), where the missing covariate Z has coefficient $C = 5$. In spite of this, the ordered values of (3.3) plot along the diagonal, as shown in Figure (3). If we knew that the data was created with a $\widehat{\Lambda}_0(t) \equiv 1$, then we may impose this constraint on the survivor functions. From

[3](3.2) are the exponentials of the Cox Snell residuals; equal up to a constant to the Martingale residual, used in the counting process approach. The Cox Snell residuals are exponentially distributed if the model is correct.

[4]The following simulations were performed with EXCEL and checked with S+.

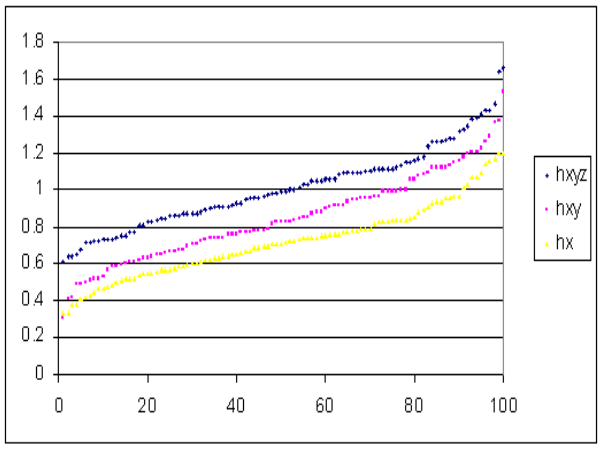

FIGURE 1. 100 ordered estimates of A for h_{XYZ}, h_{XY}, h_X, $X, Y, Z \sim U[-1, 1], (A, B, C) = (1, 1, 1)$; each estimate based on 100 samples

Figure (4) we see that uniformity is lost for models the incomplete models hy_{XY}, h_X; but not for the complete model h_{XYZ}. This would provide an excellent diagnostic for completeness if we had a priori knowledge of the baseline hazard; unfortunately in practice we do not have this knowledge. We can, however, find another diagnostic (see below).

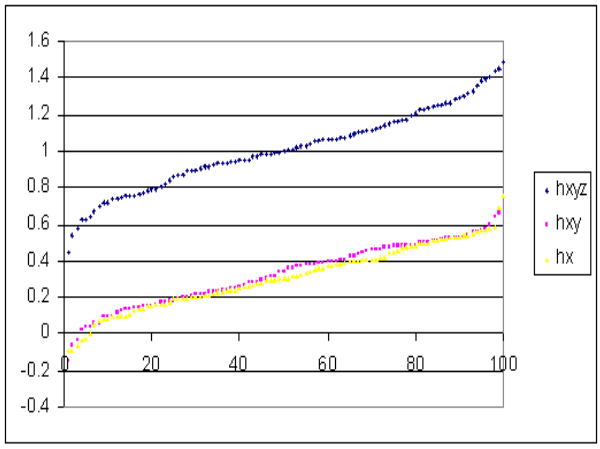

FIGURE 2. 100 ordered estimates of A for h_{XYZ}, h_{XY}, h_X, $X, Y, Z \sim U[-1, 1], (A, B, C) = (1, 1, 5)$ each estimate based on 100 samples

Figure (5) shows the Wald 95% confidence bounds for A in model h_X, in each of the 100 repetitions of the experiment whose estimates are shown in Figure (2).

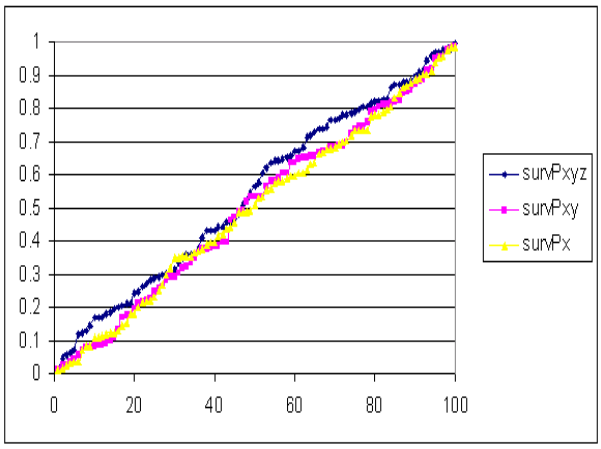

FIGURE 3. Ordered values of (3.3) for h_{XYZ}, h_{XY}, h_X, $X, Y, Z \sim U[-1,1], (A,B,C) = (1,1,5)$

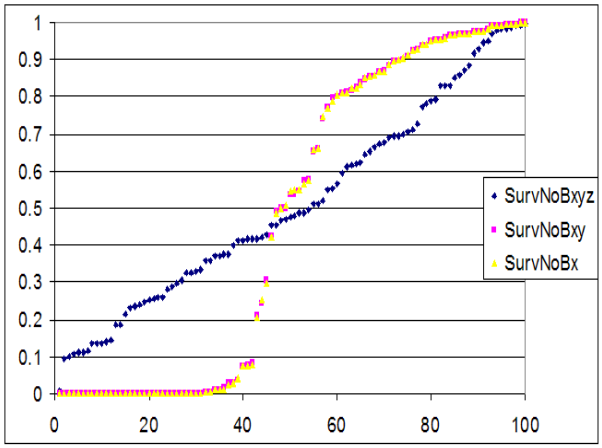

FIGURE 4. Ordered values of (3.3) for h_{XYZ}, h_{XY}, h_X, $X, Y, Z \sim U[-1,1], (A,B,C) = (1,1,5)$ with $\widehat{\Lambda}_0(t) \equiv 1$;

These bounds are derived assuming asymptotic normality of the Wald statistic

$$\frac{\hat{A} - A}{\sigma_A}$$

where $\hat{A}$ is the estimate of A and σ_A is derived from the observed information matrix. If the likelihood function is correct, then the Wald statistic is asymptotically standard normal. In as much as these 95% confidence bands contain the true value $A = 1$ in only 7% of the cases, the wisdom of stating such confidence bounds when model adequacy cannot be demonstrated may be questioned.

The models h_{XY} and h_X are clearly incorrect and mis-estimate the covariate A. Relative risk coefficients based on these models would be biased. Without

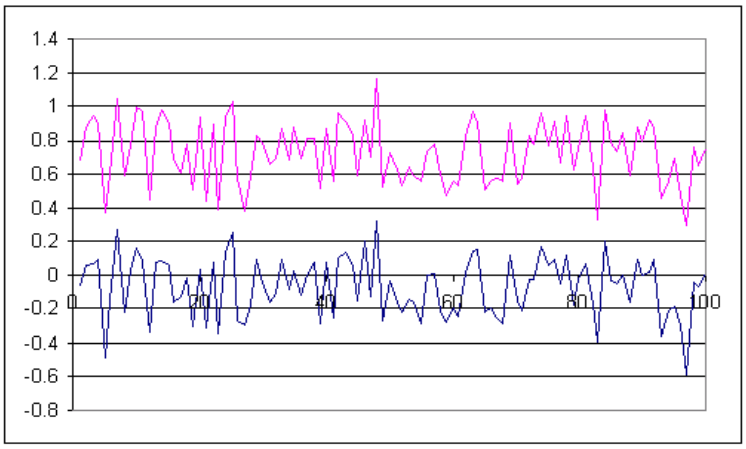

FIGURE 5. Wald 95% confidence bounds for A with model , h_X of Figure (2); each estimate based on 100 samples

a priori knowledge of the baseline hazard function, their incorrectness cannot be diagnosed using Cox-Snell or Martingale residuals, echoing the statements cited in the footnote at the beginning of this section. The problem is that the lack of fit caused by missing covariates is compensated in the estimated baseline hazard function.

This observation suggests that we might detect lack of fit in the covariates by comparing the estimated baseline hazard function with the population cumulative hazard function. From (3.3) it is evident that adding a constant to any covariate is equivalent to multiplying the baseline hazard by a constant. We therefore standardize the covariates by centering their distributions on the means (the distributions here already centered). Figures (6, 7) show these comparisons for the two cases from Figures (1,2). Note the difference in survival times (horizontal axis); this is caused by the heavier loading of covariate Z in Figure (7). The Nelson Aalen estimator is used for the population cumulative hazard function.

We see in Figure (7) that the cumulative baseline hazard functions for h_{XY} and h_X have moved closer to the population cumulative hazard , reflecting the heavier loading on the missing covariate Z.

If a Cox model had *none* of the actual covariates, this would be equivalent to having zero coefficients on all covariates; and in this case the baseline hazard would coincide with the population cumulative hazard. A simple heuristic test of model adequacy would test the null hypothesis that the cumulative baseline hazard function is equal to the population cumulative hazard function. If the null hypothesis cannot be rejected, then using the Cox model would not be indicated. In Figures (8, 9) the asymptotic 2-sigma bands on the asymptotic variance of the Nelsen Aalen estimator of the population cumulative hazard function ([16], p.25) have been added to Figures (6, 7). We see that with this simple test we would fail to reject the null hypothesis for model h_X after 100 observations in both cases. The greater loading of missing covariate Z in Figure (9) causes the model h_{XY} to fail to reject the null hypothesis as well.

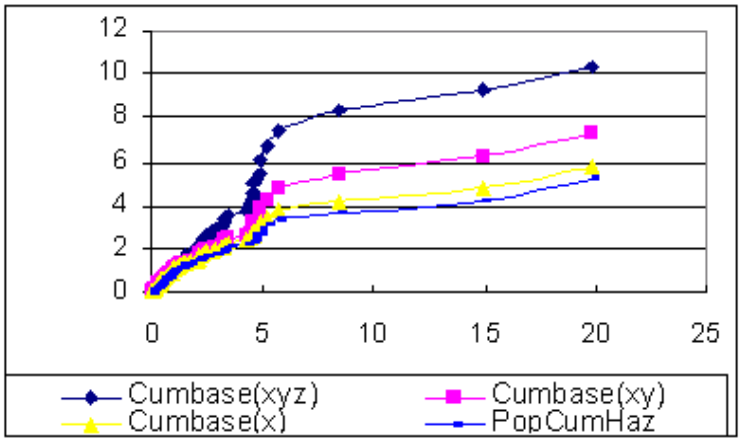

FIGURE 6. Cumulative population and baseline hazard functions for $h_{XYZ}, h_{XY}, h_X, X, Y, Z \sim U[-1,1], (A,B,C) = (1,1,1)$

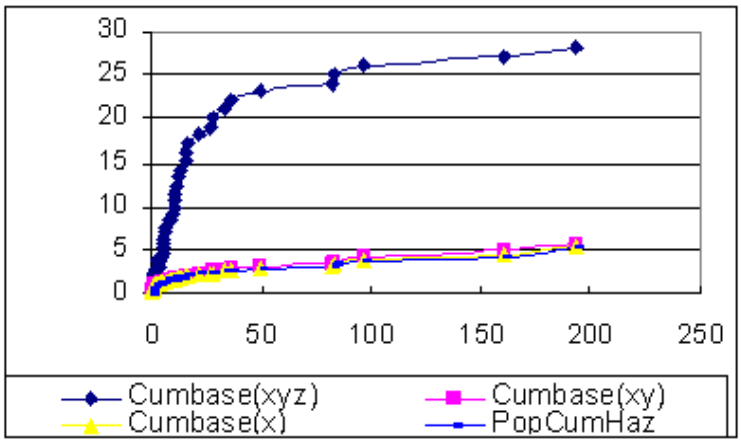

FIGURE 7. Cumulative population and baseline hazard functions for $h_{XYZ}, hy_{XY}, h_X, X, Y, Z \sim U[-1,1], (A,B,C) = (1,1,5)$

The more familiar partial likelihood ratio test calculates the test statistic G as twice the difference between the the log partial likelihood of the model containing the covariates and the log partial likelihood for the model not containing the covariates. G is asymptotically chi square distributed under the null hypothesis. The above test may have some advantage in that it does not appeal to partial likelihood. However, it is unable to detect the lack of fit in the model h_{XY} when $C = 1$.

We note that for all the results mentioned above, the covariates are independent. In practice independence is not usually checked, and not always plausible. Figure (10) shows 100 estimates of the coefficient A for the models h_{XYZ}, hy_{XY}, h_X where the covariates are uniformly distributed on $[0,1]$ with correlations $\rho(X,Z) = 0.98, \rho(Y,Z) = 0.41$ (the lack of centering has no effect on the coefficient estimates).

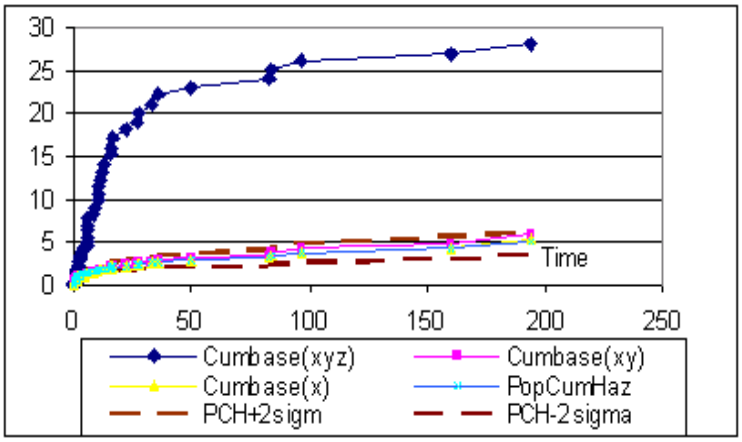

FIGURE 8. Cumulative population and baseline hazard functions for $h_{XYZ}, hy_{XY}, h_X, X, Y, Z \sim U[-1,1], (A,B,C) = (1,1,5)$ with 2-sigma confidence bands (dashed lines)

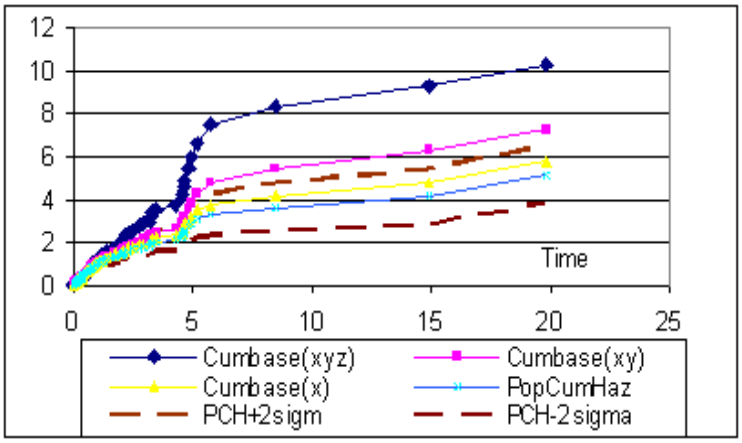

FIGURE 9. Cumulative population and baseline hazard functions for $h_{XYZ}, hy_{XY}, h_X, X, Y, Z \sim U[-1,1], (A,B,C) = (1,1,1)$ with 2-sigma confidence bands (dashed lines)

Whereas missing covariates produce under-estimation in the case of independence, we see that dependence in Figure(10) produces over-estimation. Note also that the spread of estimates for the complete model h_{XYZ} is very wide.

4. Censoring and Competing Risk

The discussion of model adequacy with the proportional hazard model is sometimes clouded by the role of censoring. The following statement is representative: "A perfectly adequate model may have what, at face value, seems like a terribly low R^2 due to a high percent of censored data" ([15] p. 229). The reference to R^2

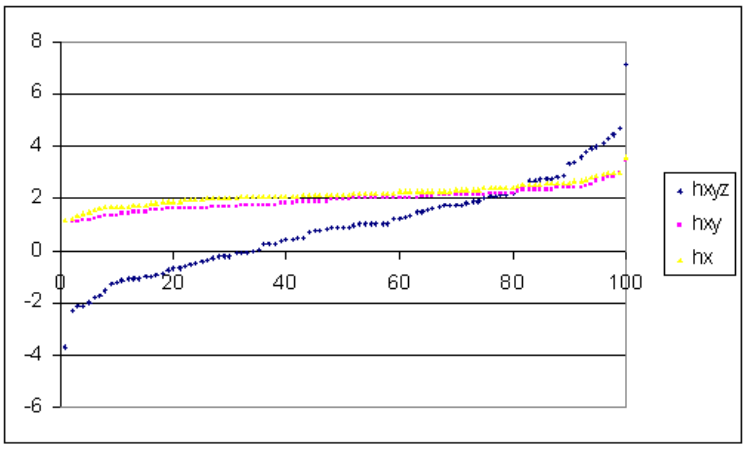

FIGURE 10. 100 ordered estimates of A for $h_{XYZ}, hy_{XY}, h_X, X, Y, Z \sim U[0,1], (A,B,C) = (1,1,1)$ with dependent covariates

must be taken as metaphorical. The proportional hazard model proposes a linear regression of the log hazard function. The hazard function is not observed, and hence a measure of the difference between observed and predicted values, like R^2 is not meaningful. The point is that the ability of a proportional hazard model to "explain the data" might be obscured by censoring.

Right censoring of course is a form of competing risk. In this section we review some recent results from the theory of competing risk, and indicate how they may yield diagnostic tools in proportional hazard modelling. In the competing risks approach we model the data as a sequence of i.i.d. pairs (T_i, δ_i), $i = 1, 2, \ldots$. Each T is the minimum of two or more variables, corresponding to the competing risks. We will assume that there are two competing risks, described by two random variables D and C such that $T = \min(D, C)$. D will be time of death which is of primary interest, while C is a censoring time corresponding to termination of observation by other causes. In addition to the time T one observes the indicator variable $\delta = I(D < C)$ which describes the cause of the termination of observation. For simplicity we assume that $P(D = C) = 0$.

It is well known (Tsiatis[23]) that from observation of (T, δ) we can identify only the subsurvivor functions of D and C:

$$\begin{aligned} S_D^*(t) &= P(D > t, D < C) = P(T > t, \delta = 1) \\ S_C^*(t) &= P(C > t, C < D) = P(T > t, \delta = 0), \end{aligned}$$

but not in general the true survivor functions of D and C, $S_D(t)$ and $S_C(t)$. Note that $S_D^*(t)$ depends on C, though this fact is suppressed in the notation. Note also that $S_D^*(0) = P(D < C) = P(\delta = 1)$ and $S_C^*(0) = P(C < D) = P(\delta = 0)$, so that $S_D^*(0) + S_C^*(0) = 1$.

The conditional subsurvivor functions are defined as the survivor functions conditioned on the occurrence of the corresponding type of event. Assuming continuity

of $S_D^*(t)$ and $S_C^*(t)$ at zero, these functions are given by

$$\begin{aligned} CS_D^*(t) &= P(D > t | D < C) = P(T > t | \delta = 1) = S_D^*(t)/S_D^*(0) \\ CS_C^*(t) &= P(C > t | C < D) = P(T > t | \delta = 0) = S_C^*(t)/S_C^*(0). \end{aligned}$$

Closely related to the notion of subsurvivor functions is the probability of censoring beyond time t,

$$\Phi(t) = P(C < D | T > t) = P(\delta = 0 | T > t) = \frac{S_C^*(t)}{S_D^*(t) + S_C^*(t)}.$$

This function has some diagnostic value, aiding us to choose among competing risk models to fit the data. Note that $\Phi(0) = P(\delta = 0) = S_C^*(0)$.

As mentioned above, without any additional assumptions on the joint distribution of D and C, it is impossible to identify the marginal survivor functions $S_D(t)$ and $S_C(t)$. However, by making extra assumptions, one may restrict to a class of models in which the survivor functions are identifiable. A classical result on competing risks [23, 24, 20] states that, assuming independence of D and C, we can determine uniquely the survivor functions of D and C from the joint distribution of (T, δ), where at most one of the survivor functions has an atom at infinity. In this case the survivor functions of D and C are said to be identifiable from the censored data (T, δ). Hence, an independent model is always consistent with data.

If the censoring is assumed to be independent then the survivor function for T, the minimum of D and C, can be written as

$$S_T(t) = S_D(t) S_C(t) \tag{4.1}$$

If we assume that D obeys a proportional hazard model, and that the censoring is independent, then we may estimate the coefficients of by maximizing the partial likelihood function adapted to account for censoring:

$$\prod_{i \in D_N} \frac{e^{x_i A + y_i B + z_i C}}{\sum_{j \geq i}^{n} e^{x_j A + y_j B + z_j C}} \tag{4.2}$$

where D_N is the subset of observed times $t_1, ...t_N$ at which death is observed to occur, and j runs over all times corresponding to death or censoring.

If we now substitute the survivor function with estimated coefficients into (4.1), and use the familiar Kaplan Meier estimator for S_C, then we may apply the ideas of the previous section to assess model adequacy.

4.1. **Independent Exponential Competing Risks.** A model in which D and C are independent is always consistent with the data, but an independent *exponential* model is not in general consistent with the data. One can derive a sharp criterion for independence and exponentiality in terms of the subsurvivor functions [7]:

Theorem 4.1. *Let D and C be independent life variables. Then any two of the following conditions imply the others:*

$$\begin{aligned} S_D(t) &= \exp(-\lambda t) \\ S_C(t) &= \exp(-\gamma t) \\ S_D^*(t) &= \frac{\lambda}{\lambda+\gamma}\exp(-(\lambda+\gamma)t) \\ S_C^*(t) &= \frac{\gamma}{\lambda+\gamma}\exp(-(\lambda+\gamma)t) \end{aligned}$$

Thus if D and C are independent exponential life variables with failure rates λ and γ, then the conditional subsurvivor functions of D and C are equal and correspond to exponential distributions with failure rate $\lambda+\gamma$. Moreover, the probability of censoring beyond time t is constant. Thus

$$\begin{aligned} CS_D^*(t) &= CS_C^*(t) = \exp(-(\lambda+\gamma)t) \\ \Phi(t) &= \frac{\gamma}{\lambda+\gamma}. \end{aligned}$$

4.2. **Random Signs Censoring.** Perhaps the simplest dependent competing risk model which leads to an identifiable marginal distribution of D is random signs censoring [7]. Suppose that the event that the time of death of a subject is censored is independent of the age D at which the subject would die, but given that the subject's time of death is censored, the time at which it is censored may depend on D[5]. This situation is captured in the following definition:

Definition 4.2. Let D and C be life variables with $C = D - W\delta$, where $0 < W < D$ is a random variable and δ is a random variable taking values $\{1, -1\}$, with D and δ independent. The variable $T \equiv [\min(D, C), I(D < C)]$ is called a random sign censoring of D by C.

Note that in this case

$$\begin{aligned} S_D^*(t) &= Pr\{D > t, \delta = -1\} = Pr\{D > t\}Pr\{\delta = -1\} = \\ &= S_D(t)Pr\{C > D\} = S_D(t)S_D^*(0). \end{aligned}$$

Hence $S_D(t) = CS^*(t)$ and it follows that the distribution of D is identifiable under random signs censoring.

A joint distribution of (D, C) which satisfies the random signs requirement, exists if and only if $C_D^*(t) > C_C^*(t)$ for all $t > 0$ [7]. In this case the probability of censoring beyond time t, $\Phi(t)$, is maximum at the origin.

4.3. **Conditional Independence Model.** Another model from which we have identifiability of marginal distributions is the conditional independence model introduced by Hokstad [14, 12]. This model considers the competing risk variables D and C to be sharing a common quantity, V, and to be independent given V. More precisely, the assumption is that

$$D = V + W, \quad C = V + U,$$

where V, U, W are mutually independent. Hokstad and Jensen[14] derived explicit expressions for the case when V, U, W are exponentially distributed:

[5]For applications of this model in reliability, see [8, 6]

Theorem 4.3. *Let V, U, W be independent with $S_V(t) = e^{-\lambda_V t}, S_U(t) = e^{-\lambda_U t}, S_W(t) = e^{-\lambda_W t}$. Then*

$$\begin{aligned} S_D^*(t) &= \frac{\lambda_V \lambda_W e^{-(\lambda_U+\lambda_W)t}}{(\lambda_U+\lambda_W)(\lambda_V-\lambda_W-\lambda_U)} - \frac{\lambda_W e^{-\lambda_V t}}{\lambda_V-\lambda_W-\lambda_U} \\ S_C^*(t) &= \frac{\lambda_V \lambda_U e^{-(\lambda_U+\lambda_W)t}}{(\lambda_U+\lambda_W)(\lambda_V-\lambda_W-\lambda_U)} - \frac{\lambda_U e^{-\lambda_V t}}{\lambda_V-\lambda_W-\lambda_U} \\ CS_D^*(t) &= CS_C^*(t) = S_D^*(t) + S_C^*(t) \\ \Phi(t) &= \frac{\lambda_U}{\lambda_U+\lambda_W} \end{aligned}$$

Moreover, if V has an arbitrary distribution such that $P(V \geq 0) = 1$, and V is independent of U and W, then still we have

$$CS_D^*(t) = CS_C^*(t)$$

Thus, as in the case of independent exponential competing risks we have equal conditional subsurvivor functions, and the probability of censoring beyond time t, $\Phi(t)$, is constant. However, the conditional subsurvivor functions need not be exponential. Nothing is know about their general form.

4.4. **Mixture of Exponentials Model.** Suppose that $S_D(t)$ is a mixture of two exponential distributions with parameters λ_1, λ_2 and mixing coefficient p, and that the censoring survivor distribution $S_C(t)$ is exponential with parameter λ_y:

$$\begin{aligned} S_D(t) &= p\exp\{-\lambda_1 t\} + (1-p)\exp\{-\lambda_2 t\} \\ S_C(t) &= \exp\{-\lambda_y t\}. \end{aligned}$$

The properties of the corresponding competing risk model is given by [5].

Theorem 4.4. *Let D and C be independent life variables with the above distributions. Then,*

$$\begin{aligned} S_D^*(t) &= p\frac{\lambda_1}{\lambda_y+\lambda_1}\exp\{-(\lambda_y+\lambda_1)t\} + (1-p)\frac{\lambda_2}{\lambda_y+\lambda_2}\exp\{-(\lambda_y+\lambda_2)t\} \\ S_C^*(t) &= p\frac{\lambda_y}{\lambda_y+\lambda_1}\exp\{-(\lambda_y+\lambda_1)t\} + (1-p)\frac{\lambda_y}{\lambda_y+\lambda_2}\exp\{-(\lambda_y+\lambda_2)t\} \\ CS_D^*(t) &= \frac{\left(\exp\{-(\lambda_y+\lambda_1)t\} + \frac{1-p}{p}\frac{\lambda_2}{\lambda_1}\frac{\lambda_y+\lambda_1}{\lambda_y+\lambda_2}\exp\{-(\lambda_y+\lambda_2)t\}\right)}{\left(1+\frac{1-p}{p}\frac{\lambda_2}{\lambda_1}\frac{\lambda_y+\lambda_1}{\lambda_y+\lambda_2}\right)} \\ CS_C^*(t) &= \frac{\left(\exp\{-(\lambda_y+\lambda_1)t\} + \frac{1-p}{p}\frac{\lambda_y+\lambda_1}{\lambda_y+\lambda_2}\exp\{-(\lambda_y+\lambda_2)t\}\right)}{\left(1+\frac{1-p}{p}\frac{\lambda_y+\lambda_1}{\lambda_y+\lambda_2}\right)} \\ CS_D^*(t) &\leq CS_C^*(t) \end{aligned}$$

Moreover, $\Phi(t)$ is minimal at the origin, and is strictly increasing when $\lambda_1 \neq \lambda_2$.

4.5. **Heuristics for Model Selection.** The probability $\Phi(t)$ of censoring after time t, yields a diagnostic for model selection, together with the conditional subsurvivor functions $CS_D^*(t)$ and $CS_C^*(t)$. Statistical tests are developed in [4]. The following statements, which follow from the results of the previous subsections, may guide in model selection.

- If the risks are exponential and independent, then the conditional subsurvivor functions are equal and exponential. Moreover, $\Phi(t)$ is constant.
- Under random signs censoring, $\Phi(0) > \Phi(t)$ and $CS^*_D(t) > CS^*_C(t)$ for all $t > 0$.
- If the conditional independence model holds with U, W exponential, then the conditional subsurvivor functions are equal and $\Phi(t)$ is constant
- If the mixture of exponentials model holds, then $\Phi(t)$ is strictly increasing and $CS^*_D(t) \leq CS^*_C(t)$ for all $t > 0$.

5. Example

We illustrate the ideas with a data set on lung cancer patients from the Mayo Clinic [18]. The data involve 165 observed times of death and 63 censoring times, 228 times in total. The censoring is assumed to be independent. 8 covariates are used to construct a proportional hazard model.

We first obtain the coefficient values which maximize the partial likelihood (4.2). We then estimate the baseline hazard at each observed time of death, as described in [16][6]. We see that the cumulative baseline hazard is nearly linear up to 883 days, indicating a nearly constant baseline hazard rate. The last observations are censors; the fact that the baseline hazard rate is estimated only at times of death explains the flat shape after $t = 883$.

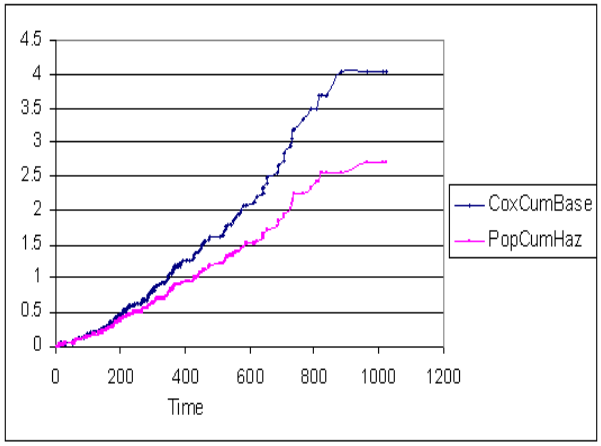

Figure 11. Cumulative baseline hazard and population cumulative hazard for Mayo clinic lung cancer data

Figure (11) shows the Cox cumulative baseline hazard function and the population cumulative hazard function. Figure (12) adds the 2-sigma bounds from the asymptotic variance of the Nelson Aalen estimate. The Cox baseline hazard function nearly coincides with the upper 2-sigma curve. Figure (13) shows the conditional subsurvivor functions for death and censoring, and shows the function $\Phi(t)$. Note that the conditional subsurvival function for censoring dominates that

[6]There are a few ties in this data set which would significantly complicate the calculations of the baseline hazard. We therefore broke the ties by adding small increments, verifying that this had negligible effect on the results.

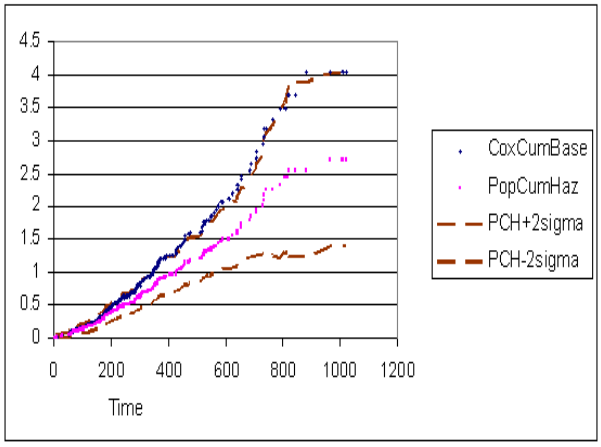

FIGURE 12. Cumulative baseline hazard and population cumulative hazard for Mayo clinic lung cancer data with 2 sigma confidence bands

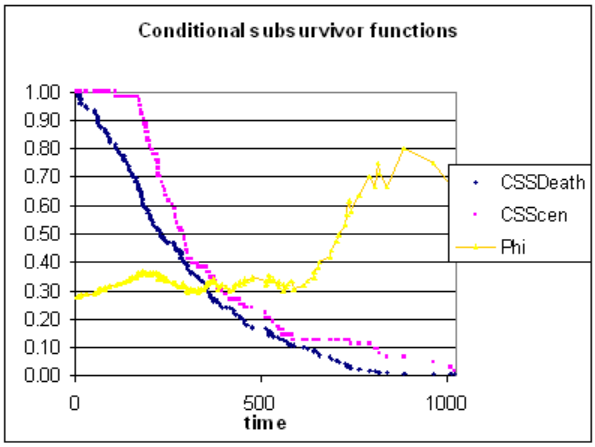

FIGURE 13. Conditional subsurvivor functions, and $\Phi(t)$ for Mayo clinic Lung cancer data

for death, and the $\Phi(t)$ function is roughly increasing, up to the time of the last observed death (883), after which the conditional subsurvivor for death is constant and $\Phi(t)$ therefore decreases. This is the pattern we should expect if a mixture of exponential life variables is censored independently by an exponential variable.

The picture which emerges is mixed. On the one hand the Cox model with constant covariates is barely able to distinguish the cumulative baseline hazard and population cumulative hazard functions. On the other hand, the conditional subsurvivor functions are consistent with independent censoring of a mixture of exponentials with an exponential censoring variable. If this censoring mechanism

were *not* true, then we should have to come up with another explanation for the distinctive pattern in Figure(13). Taken together, these considerations would motivate finding other covariates to add to the Cox model.

6. CONCLUSION

Subsurvivor diagnostics can help us recognize censoring patterns associated with certain types of dependent censoring and/or certain classes of life distributions. The Cox proportional hazard with constant covariates entails a mixed exponential live distribution.

REFERENCES

[1] Aras,G. and Deshpande, J.V. (1992) Statistical analysis of dependent competing risks, *Statistics and Decisions* 10, 323-336.

[2] Allison, P.D. (2003) "Survival Analysis Using SAS a practical guide" SAS Publishing, Cary.

[3] Bretagnolle, J. and Huber-Carol, C. (1988)Effects of omitting covariates in Cox's Model for susrvival data"Scand. J. Statist. 15, 125-138.

[4] C. Bunea, R.M. Cooke and B. Lindqvist, Analysis tools for competing risk failure data, European Network for Business and Industrial Statistics Conference, ENBIS 2002, Rimini, Italy, 23-24 September 2002.

[5] C. Bunea, R.M. Cooke, and B. Lindqvist "Competing risk perspective over reliability data bases" Mathematical and Statistical Methods in Relibiity (B.H. Lindqvist and K.A.Doksum eds), World Scientific Publishing, Singapore, 2003,p 355-370.

[6] C. Bunea, R.M. Cooke and B. Lindqvist, Maintenance study for components under competing risks, in *Safety and Reliability, 1st volume*, (European Safety and Reliability Conference, ESREL 2002, Lyon, France, 18-21 March 2002), pp. 212-217.

[7] R.M. Cooke, The design of reliability databases Part I and II, *Reliability Engineering and System Safety*, **51**, 137-146 and 209-223 (1996).

[8] R.M. Cooke and T.J. Bedford, Reliability databases in perspective, *IEEE Transactions on Reliability*, vol. 51, no 3, (September 2002), pp. 294-310.

[9] R.M. Cooke, T.J. Bedford, I. Meilijson and L. Meester, Design of reliability databases for aerospace applications, Report to the European Space agency, Department of Mathematics Report 93-110, Delft University of Technology, 1993.

[10] D.R. Cox, 1972. "Regression Models and Life-Tables" Royal Statistical Society, Series B, Vol. 34, No. 2, 1972, pp. 187-220.

[11] Dockery, D., III, C. A. P., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G., and Speizer, F. E. 1993. "An association between air pollution and mortality in six U.S. cities." New England Journal of Medicine, 329, 1753-1759.

[12] J. Dorrepaal, P. Hokstad, R.M. Cooke and J.L. Paulsen, The effect of preventive maintenance on component reliability, in *Advances in Safety and Reliability*, Ed. C.G. Soares (Proceedings of the ESREL '97 conference, 1997), pp. 1775-1781.

[13] S.E. Erlingsen, Using reliability data for optimizing maintenance, unpublished master's thesis, Norwegian Institute of Technology, 1989.

[14] P. Hokstad and R. Jensen, Predicting the failure rate for components that go through a degradation state, *Reliability Engineering and System Safety,* **53**, 389-396 (1998).

[15] Hosmer, D.W. and Lemeshow, S. (1999) Applied Survival Analysis, Wiley, New York

[16] Kalbfleisch, J.D. and Prentice, R.L. (2002) The Statistical Analysis of Failure Time Data, second edition. Wiley New York.

[17] Keiding, N., Andersen, P.K., Klein, J.P. (1997) "The role of frailty time models in describing heterogeneity due to omitted covariates"*Statistic's in Medicine*, vol. 16, 215-224.

[18] Loprinzi, C.L. Goldberg, R.M. Su, J.Q. Mailliard, J.A. Kuross, S.A. Maksymiuk, A.W. Kugler, J.W. Jett, J.R. Ghosh, C. Pfeille, D.M. and Burch; I (1994) Placebo-controlled trial of htydrazine sulfate in patients with newly diagnosed non-small-cell lung cancer" J. of Clinical Oncology, Vol. 12, No. 6 1126-1129.

[19] Oakes, D. (2001) "Biometrika Centenary: Survival analysis" Biometrika 88 99-142.

[20] A.V. Peterson, Bounds for a joint distribution function with fixed subdistribution functions: Application to competing risks, *Proceedings of the National Academy of Sciences, USA,* **73**, 11-13 (1976).
[21] Pope, C. A., Thun, M. J., Namboodiri, M. M., Dockery, D., Evans, J. S., Speizer, F. E., and Heath, C. W. 1995. "Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults." American Journal of Respiratory and Critical Care Medicine, 151, 669-674.
[22] Therneau, T.M. and Grambsch, P.M. (2000) Modeling Survival Data, Springer, New York.
[23] A. Tsiatis, A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Sciences, USA*, **72**, 20-22 (1975).
[24] van der Weide, J.A.M. van der, and Bedford T., "Competing risks and eternal life", in *Safety and Reliability (Proceedings of ESREL'98)*, S. Lydersen, G.K. Hansen, H.A. Sandtorv (eds), Vol 2, 1359-1364, Balkema, Rotterdam, 1998.

Delft University of Technology, Department of Mathematics, Mekelweg 4, NL-2628 CD Delft, The Netherlands
E-mail address: r.m.cooke@its.tudelft.nl

Delft University of Technology, Department of Mathematics, Mekelweg 4, NL-2628 CD Delft, The Netherlands
E-mail address: o.moralesnapoles@ewi.tudelft.nl