A Graphical Method for Assessing Goodness of Fit in Cox's Proportional Hazards Model

ELJA ARJAS*

Suggested here is a simple graphical method for studying the goodness of fit in Cox's regression model for survival data. The method is easy to use, as it does not require the estimation of alternative models and only involves quantities similar to those already appearing in the partial likelihood expression that is needed in the parameter estimation. The rationale behind the graphs is very intuitive: They make a direct comparison between observed and expected failure frequencies, as estimated from the model. In a correctly specified model one anticipates an approximate balance between such frequencies; otherwise there will typically be groups of individuals for which the expected frequencies are systematically too high or too low to match with the data, and this shows in the graphs introduced here. In the concrete applications of the method the individuals are stratified in a way that depends on what aspect of the model is being checked against data. There is always one graph for each stratum. Simulated and real data are used to illustrate the method. In the simulations two types of defect that can come up in a Cox's model are considered: (a) an influential covariate has been deleted from the model, and (b) a common baseline hazard for all individuals has been assumed in a case in which the individuals should be stratified according to baseline hazard. Serious defects in the model are relatively easy to detect from the diagnostic graphs. As concrete applications of the method, studied briefly are the fitting of Cox's model to the well-known Stanford heart transplant data and to a data set describing the survival of malignant melanoma patients after operation. The article concludes with some general observations concerning the randomness in the graphs.

1. INTRODUCTION AND PRELIMINARIES

With the growing popularity of the semiparametric proportional hazards model of Cox (1972), it has become increasingly important to find convenient ways to detect when such a model is poorly specified. My impression is that few articles reporting on the application of Cox's model on survival data actually perform a goodness-of-fit analysis.

The literature on goodness-of-fit techniques suitable for Cox's model is currently fairly extensive. One should first note that the simple two-sample model assuming constant proportionality between two otherwise unspecified hazard rates is a special case of Cox's model. There are well-known methods based on total time (see, e.g., Aalen and Hoem 1978; Barlow and Campo 1975; Gill and Schumacher 1987) suitable for detecting when the proportionality assumption is not met. For the more general regression model using covariates, for example, Kay (1977), Crowley and Hu (1977), and Crowley and Storer (1983) used cross-plots of estimated "generalized residuals," either against a covariate value or against a set of order statistics from the unit exponential distribution. Lagakos (1981) recommended a method based on permuted rank statistics of such residuals. Andersen (1982) studied graphical techniques similar to those in Kay (1977) and derived a goodness-of-fit test that involves the estimation of a piecewise constant baseline hazard. Cox (1979) introduced a test statistic based on a stratification and the corresponding operational time. Two recent papers apply weak convergence theory: Wei (1984) considered the constant proportionality assumption in a two-sample case, establishing the Brownian bridge as the asymptotic limit of the score process, and Hjort (1984) gave the baseline hazard a parametric form and then studied the asymptotic behavior of the integrated hazards. Wei's result was generalized in Haara (1987). For other methods of assessing goodness of fit in Cox's proportional hazards model, see Schoenfeld (1980, 1982), Lagakos and Schoenfeld (1984), Breslow, Edler, and Berger (1984), and Moreau, O'Quigley, and Mesbah (1985).

Let us first recall the structure of the proportional hazards model in its original form. Consider a set of n individuals, indexed by j ($1 \le j \le n$). With each individual we associate an observation (T_j, δ_j), where T_j is the failure time or the censoring time, whichever occurs first, and $\delta_j = 1$ or 0 depending on whether j fails or is censored at T_j. The proportional hazards model of Cox (1972) postulated that the failure intensity of j at time t can be expressed in the form

$$\lambda_j(t) = Y_j(t) \cdot \lambda_0(t)e^{\beta z_j}, \qquad (1.1)$$

where $Y_j(t) = 1_{\{T_j \ge t\}}$ is the indicator of being at risk at time t, $\lambda_0(t)$ is an unspecified baseline hazard, β is a p-vector of model parameters, and z_j is a p-vector of (fixed) covariates describing individual j. In this simple setting tied failure times are ruled out. The variable t usually measures the distance from an entry time (which can depend on j).

Let P^β be a probability that is in agreement with the foregoing assumptions and write $P = P^{\beta_0}$, where β_0 is the "true" parameter value.

I follow Andersen and Gill (1982) in extending this setting to allow for time-dependent (and possibly random)

* Elja Arjas is Professor, Department of Applied Mathematics and Statistics, University of Oulu, 90570 Oulu, Finland. A first version of this work was done while the author was visiting the Fred Hutchinson Cancer Research Center, Seattle, Washington. It was supported in part by National Institutes of Health Grants 5R01 GM-28314, 7R01 CA-39949, and 5R01 GM-24472. The computations for Sections 3 and 4 were done by David Venzon and Pekka Kangas. Their skillful aid was crucial to the completion of this work and is gratefully acknowledged. The author is also grateful to Søren Johansen, Per Andersen, John Crowley, Edward Lustbader, and the referees, for comments, and to K. T. Drzewiecki for permission to use the malignant melanoma data.

covariates. Denote by $N_j(t) = 1_{\{T_j \le t, \delta_j = 1\}}$ the counting process that counts "one" when the jth individual fails (uncensored), and let $\bar{N}(t) = \sum_{j=1}^n N_j(t)$. Denote by $Z_j(t)$ the covariate p-vector for individual j at time t. It is then assumed that, under P^β, the proportional hazards model (1.1) holds in the more general form $\lambda_j(t) = Y_j(t) \cdot \lambda_0(t)e^{\beta Z_j(t)}$. Here the intensity $\lambda_j(t)$ has the interpretation $\lambda_j(t)\,dt = P^\beta(T_j \in dt, \delta_j = 1 \mid \mathcal{F}_{t-})$, $\mathcal{F}_t$ being the history describing the pre-t information about survival, censoring, and covariates. For technical reasons, assume that $t \to Z_j(t)$ is left-continuous for $t \le T_j$, and for $t > T_j$ set $Z_j(t) = 0$.

Note that the usual structure $e^{\beta Z}$ of the relative risk function is assumed here for convenience only. The diagnostic method will not depend on any particular functional form of the relative risks.

The following shorthand notation will be used:

$$p_j^\beta(t) = \frac{Y_j(t)e^{\beta Z_j(t)}}{\sum_{k=1}^n Y_k(t)e^{\beta Z_k(t)}}, \qquad t \ge 0;\ j = 1, 2, \ldots, n. \tag{1.2}$$

If $K = \sum_{j=1}^n \delta_j$ is the number of uncensored failures and $T_{(1)} < T_{(2)} < \cdots < T_{(K)}$ are the corresponding failure times, the partial likelihood L^β of Cox (1972) becomes the product of $p_j^\beta(T_{(i)})$, where i and j are such that $\Delta N_j(T_{(i)}) = 1$; that is, j is not censored and fails at $T_{(i)}$. It is also well known [compare Andersen and Gill (1982) or Gill (1984)] that the process

$$X_j^\beta(t) = N_j(t) - \int_0^t p_j^\beta(s)\, d\bar{N}(s), \qquad t \ge 0, \tag{1.3}$$

is a $(P^\beta, \mathcal{F}_t)$ martingale for each j. By considering $X_j^\beta(t)$ at the failure times $T_{(1)} < T_{(2)} < \cdots < T_{(K)}$ and denoting

$$M_j^\beta(k) \underset{\text{def}}{=} X_j^\beta(T_{(k)}) = N_j(T_{(k)}) - \sum_{i \le k} p_j^\beta(T_{(i)}), \qquad 1 \le k \le K, \tag{1.4}$$

we obtain an "imbedded" martingale of a discrete time parameter k. [More directly, the quantities $\Delta M_j^\beta(k) = M_j^\beta(k) - M_j^\beta(k-1)$ are seen to be martingale differences; see Arjas and Haara (1987) for details.] Thus

$$H_j^\beta(k) = \sum_{i \le k} p_j^\beta(T_{(i)}) \tag{1.5}$$

is the compensator that balances against the failure count $N_j(T_{(k)})$. By a slight abuse of notation, we will henceforth write $N_j(k)$ and $p_j(k)$ instead of $N_j(T_{(k)})$ and $p_j(T_{(k)})$.

The variables (1.4) give rise immediately to a large family of martingales. Later, simple summations over fixed subsets of individuals, obtained by a stratification, will be considered. Another well-known example of such a martingale is the score process (whose value at k corresponds to the partial likelihood arising from the k first failures).

My final comment in this section concerns the time variable t. Unless all individuals enter the study at the same calendar time, they usually need to be aligned to match for t, the argument of the common baseline hazard. The alignment, if necessary, obviously changes the meaning of the generated history $(\mathcal{F}_t)$ from what it would be if calendar time were used. [See Arjas (1985) for a discussion.] Luckily, the change appears to be harmless in most concrete situations. This is the case, for example, if all individuals, including their covariate evolution and possible censoring, are independent. In this article it will simply be assumed that the martingale property of (1.3) holds for the considered $(\mathcal{F}_t)$. In particular, as long as this remains valid any censoring mechanism is allowed.

2. A GRAPHICAL PLOTTING TECHNIQUE

I now explain how the actual plots are drawn. Suppose that $\{1, 2, \ldots, n\}$, the set of all individuals in the study, has in some way been stratified into s strata, say I_1, I_2, $\ldots$, I_s. The stratification usually arises in a natural way from some particular question regarding the model one wants to consider. In Section 3 two examples are considered. In the first example we study the effect of the waiting time on the posttransplant survival in the Stanford heart transplant study; then we simply divide the patients into those with waiting time at most 20 days and those with waiting time longer than 20 days. In the second example, taken from Andersen (1982), the stratification corresponds to the level of invasion in a study concerning the survival after surgery of 205 malignant melanoma patients. Once the stratification is done, there will be one graph for each stratum. In the following I is used as a generic symbol for any one of the sets $I_1, I_2, \ldots, I_s$.

The rationale of the plots is to use the "no trend" property of martingales. Clear upward or downward trends are then viewed as indications of poor model fit.

Although each $M_j^{\beta_0}$ $(1 \le j \le n)$ is a martingale under the correct model P, the individual graphs $k \to M_j^{\beta_0}(k)$ will all be first down, unless j happens to fail at $T_{(1)}$, as long as j remains at risk, finally ending with an upward jump at failure if j is not censored. These basic graphs are now modified, needing four steps.

First, the summation is done over all $j \in I$. From now on, to simplify the notation, the subscript I means summation over $j \in I$. Thus we consider $M_I^{\beta_0} = N_I - H_I^{\beta_0} = \sum_{j \in I} M_j^{\beta_0}$.The martingale property of $M_I^{\beta_0}$ is an expression of collective balance in I between the actual failures and a corresponding cumulative hazard.

Second, $M_I^{\beta_0}$ is considered only at times at which there is a failure in I, thus avoiding "toothed" graphs. Let $K(I) = \sum_{j \in I} \delta_j$ be the total number of observed failures in I. Denoting by γ_k^I the index i such that the kth failure in I occurs at $T_{(i)}$, we have $N_I(\gamma_k^I) = k$. We consider, therefore, $k \to M_I^{\beta_0}(\gamma_k^I) = k - H_I^{\beta_0}(\gamma_k^I)$ $(1 \le k \le K(I))$.

As a third step, we must take into account the fact that β_0 is unknown. Therefore, we replace β_0 in the plots by the partial maximum likelihood estimator $\hat{\beta}$. Unfortunately, the exact martingale property for the correct P then no longer holds. The quantities $H_I^\beta(k)$, however, depend continuously on β, so that, for consistent estimation of β_0, the "no trend" property of martingales will hold as an approximation. For more comments on this see Section 4.

Finally, as a fourth step we choose to plot $k \to H_I^{\hat{\beta}}(\gamma_k^I)$ $(1 \le k \le K(I))$ instead of $k \to k - H_I^{\hat{\beta}}(\gamma_k^I)$. The difference $k - H_I^{\hat{\beta}}(\gamma_k^I)$ then becomes the vertical distance of the point $(k, H_I^{\hat{\beta}}(\gamma_k^I))$ from the diagonal $y = x$. Again, recalling that $N_I(\gamma_k^I) = k$, we can view such differences as expressions of a balance between the actual count of failures in stratum I and a corresponding estimated collective cumulative hazard. Discrepancies between the statistical model and data can disturb such a balance, often leading to hazard estimates that are systematically too low or too high.

In the absence of ties, there is always one "point" for each (uncensored) failure, with x-coordinates equally spaced. In a sense, therefore, all points in the graph "look equally important," which is a useful property in the visual assessment of goodness of fit. Another is that if the proportional hazards model fits the data well, one can expect to get graphs having a particularly simple form: they are approximately linear with slope close to one. Moreover, different forms of misspecification in the model tend to lead to recognizably different shapes in the graphs. This is a useful property in the diagnostics.

3. EXAMPLES

Here we consider ways in which the graphs can be used to detect lack of fit in a proportional hazards model. Two particular ways of misspecifying a model are used as illustrations: an omitted covariate and nonproportional baseline hazards in different subgroups.

Apart from these, other structural questions could be addressed, such as stratum–covariate interactions [compare Thall and Lachin (1986)] or multiplicativity versus additivity of the relative risk function. In each case one needs to decide on a proper stratification of $\{1, 2, \ldots, n\}$ into the subsets $I_1, I_2, \ldots, I_s$; otherwise the plotting proceeds in the same way.

3.1 Omission of a Covariate

Suppose that, instead of p covariates, one should have used a $(p + 1)$-covariate model. Suppose for simplicity that this $(p + 1)$st covariate is fixed for each individual and that it appears on s different levels, say $a_1, a_2, \ldots, a_s$. If individual j has the omitted covariate on level a_r $(1 \le r \le s)$, the failure intensity $\lambda_j^{\beta_0}(t)$ is misspecified by the factor $\exp(-\beta_{0,s+1} \cdot a_r)$. Depending on the sign of the product $\beta_{0,s+1} \cdot a_r$, the resulting intensity is either too large or too small. On the other hand, as long as the risk set at a failure time $T_{(k)}$ contains a relatively large number of individuals with covariates on many levels, it is likely that, although the model has been specified incorrectly, such differences are approximately "averaged out" in the sum $\sum_j Y_j(T_{(k)}) e^{\beta_0 Z_j(T_{(k)})}$. By considering the stratification $I_r = \{$individuals for which the omitted covariate is equal to $a_r\}$ $(1 \le r \le s)$ [compare Andersen (1982); Thomas (1983)], the cumulative relative hazards in each group will all be

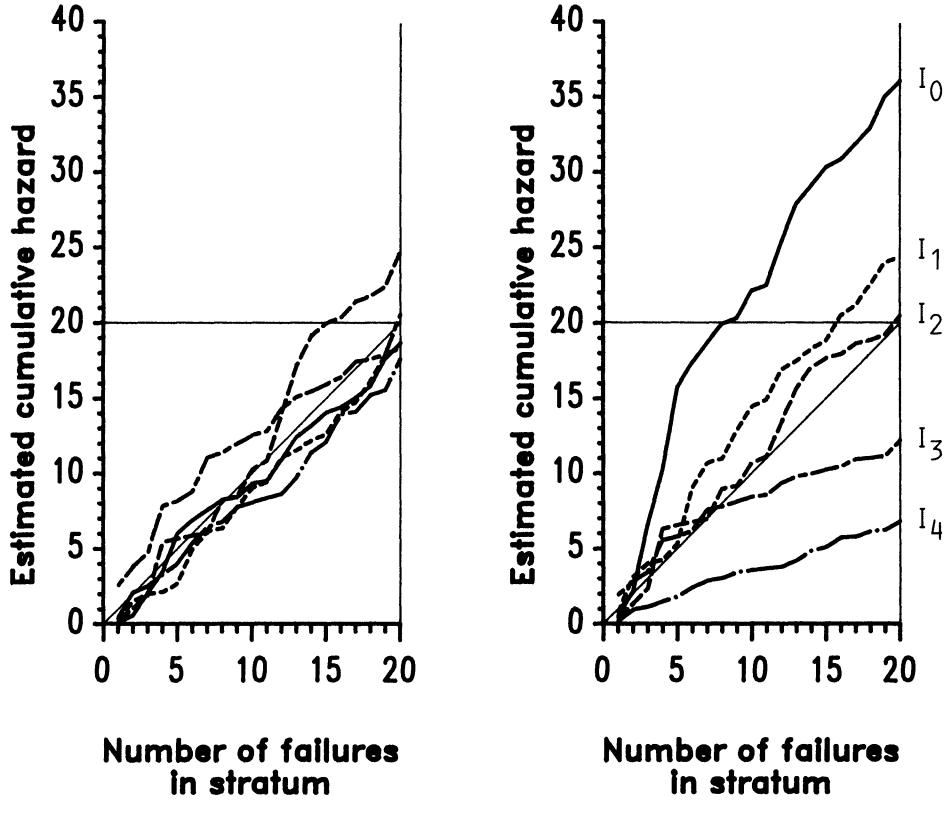

Figure 1. The Effect of Omitting an Influential Covariate (first simulation). The diagnostic curves on the left arise from fitting the correct two-covariate model, whereas on the right the second covariate was not included. The corresponding estimated regression coefficients were $\hat{\beta} = (2.18, .51)$ *(left) and* $\hat{\beta}_1 = 1.85$ *(right).*

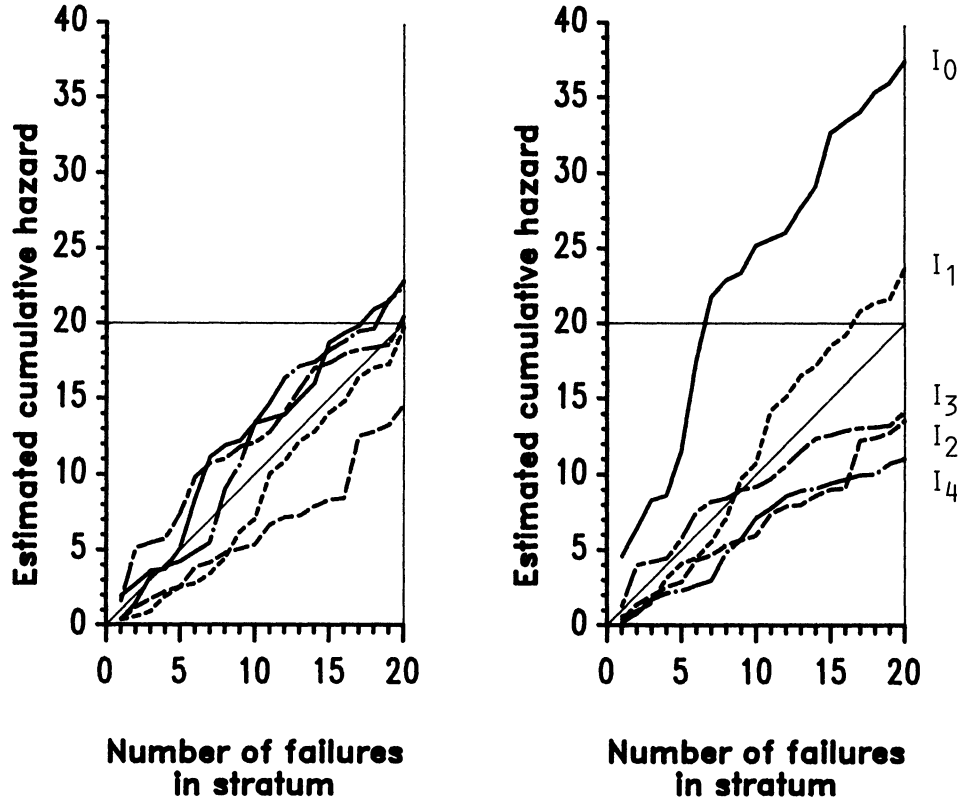

Figure 2. The Effect of Omitting an Influential Covariate (same model as in Fig. 1, second simulation). Here $\hat{\beta}$ = (2.20, .36) (left), $\hat{\beta}_1$ = 2.00 (right).

off by approximately the same factor. Thus the plots can be expected to be approximately linear. This will still hold if $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_s)$ is close to β_0 and the former is used as the parameter value.

It is interesting to compare the plots arising here to the plot "log $\hat{\Lambda}_s(t)$ versus t" in Andersen (1982). The main difference is that Andersen, and Kay (1977), compare a set of estimated cumulative baseline hazards (which are estimated separately for each stratum), whereas here, for each stratum, a direct comparison is made between an estimated cumulative hazard and the corresponding observed failure count. In this case only a single unstratified estimation is necessary. Another difference is that here we are plotting against k, instead of $T_{(k)}$. Finally, we are not using logarithms on either axis (although, of course, this is purely an option).

This diagnostic technique is illustrated in Figures 1 and 2, which give the results of two computer simulations, with $n = 100$. The same generated survival times were used in both parts of these figures. The structure of the simulation model was to have $\lambda_0(t) = 1$ and to use fixed covariates: $(Z_{j1}(t), Z_{j2}(t)) = (z_{j1}, z_{j2})$ with $z_{j1} = (j - 1)(n - 1)^{-1}$ $(1 \leq j \leq n)$ and z_{j2} taking cyclically values 0, 1, 2, 3, and 4. The corresponding coefficients were $\beta_0 = (\beta_{01}, \beta_{02}) = (2, .5)$. Corresponding to the five levels of Z_{j2} the five strata $I_r = \{1 \leq j \leq n: Z_{j2} = r\}$ $(r = 0, 1, \ldots, 4)$, each containing 20 individuals, were then considered. In both figures, the graphs on the left are based on the correct model but using the estimated $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$. In the graphs on the right the second covariate was omitted and only β_1 was estimated. There the five levels of the omitted second covariate are separated quite clearly. As expected, the curve corresponding to the smallest value of $\beta_{02} \cdot a_r$ has the highest slope, and conversely. On the left, on the other hand, the curves do not seem to be ordered in any systematic way. Other simulations (not shown here) had similar behavior.

Note that the curves on the right in Figures 1 and 2 are slightly concave. This can be explained in the following way: Typical long-time survivors have low hazard rates. If a covariate is omitted, some of this effect goes unaccounted and, consequently, the term $\Sigma_k Y_k(T_{(i)})\exp(\hat{\beta}_1 z_{k1})$, which appears as the denominator in (1.2), is biased upward when $T_{(i)}$ is large. This is reflected in low estimated values of the hazards. The effect is more pronounced in the plots corresponding to small values of $\beta_{02} \cdot a_r$, since they typically describe long-time survivors.

3.2 Falsely Assuming a Common Baseline Hazard

Sometimes a common baseline hazard $\lambda_0(t)$ is assumed in a situation where correct fitting would require the identification of s subsamples, say with baseline hazards $\lambda_{01}(t)$, $\lambda_{02}(t), \ldots, \lambda_{0s}(t)$ (Kalbfleisch 1974). If these baseline hazards are proportional we are back to the aforementioned "omitted covariate" case. If they are not, several modes of behavior may occur.

Here we consider the case of crossing baseline hazards,

assuming that the ratios $\lambda_{0r_1}(t)/\lambda_{0r_2}(t)$ are increasing in t for $r_1 < r_2$, changing from values less than 1 to values greater than 1. Now, choosing I_r to be the set $I_r = \{$individuals with baseline hazard $\lambda_{0r}(t)\}$ $(r = 1, 2, \ldots, s)$ and considering $j \in I_r$, the correct definition of the relative hazard $p_j^{\beta}(T_{(i)})$ in (1.2) would be

$$\frac{Y_j(T_{(i)})\lambda_{0r}(T_{(i)})e^{\beta Z_j(T_{(i)})}}{\sum_{q=1}^{s}\sum_{k\in I_q} Y_k(T_{(i)})\lambda_{0q}(T_{(i)})e^{\beta Z_k(T_{(i)})}}.$$

Thus the relative hazards in group I_1 are likely to be first too large when compared with the actual failure count in that group, and then too small, whereas in group I_s they are typically first too small and then too large. [Unless $n \gg$ card (I_s), the behavior may again change for the longest survival times. This is because in a sample of decreasing-failure-rate life lengths there are typically both very small and very large values and, therefore, it can be that in the last risk sets $R(T_{(k)})$ most individuals are actually long-time survivors from group I_s. In such a case, a plot for stratum I_s ultimately has a slope that is near to 1.]

Figure 3 illustrates the use of diagnostic graphs for detecting nonproportional baseline hazards. Five strata were used, with respective baseline hazards $\lambda_{01}(t) = t$, $\lambda_{02}(t) = t^{1/2}$, $\lambda_{03}(t) = 1$, $\lambda_{04}(t) = t^{-1/3}$, and $\lambda_{05}(t) = t^{-1/2}$. There were $n = 100$ individuals, evenly divided into the five groups, and a single covariate with values and true regression coefficient exactly as $Z_{j1}(t)$ and β_{01} in Figures 1 and 2. The graphs have the form anticipated previously, and this was very consistent in the simulations made.

3.3 Examples Based on Real Data

Here we consider two applications of the method on real data.

As a first illustration, we study briefly the posttransplant survival of the 65 transplanted patients in the well-known Stanford Heart Transplant data, with 41 recorded deaths. In particular, we consider the effect of the logarithmic waiting time on posttransplant survival in a Cox model, where age, mismatch score, and previous surgery are used as covariates. This question was discussed, for example, in Crowley and Storer (1983); they found that a cross-plot of generalized residuals against logarithmic waiting time revealed very little. In Figure 4 two levels of waiting time are used (group 1: up to 20 days; group 2: longer than 20 days), drawing the plots as in Section 3.1. Of course, more levels could be used in the stratification, but then the groups would become rather small for a reliable comparison. Although it appears that logarithmic waiting time should be included in the model, there is also some indication that a single proportional hazards model, with a fixed waiting time effect, does not describe such dependence well: Patients with a short waiting time seem to face a greater early risk than those who had a longer waiting time. For comparison, for this data set the corresponding

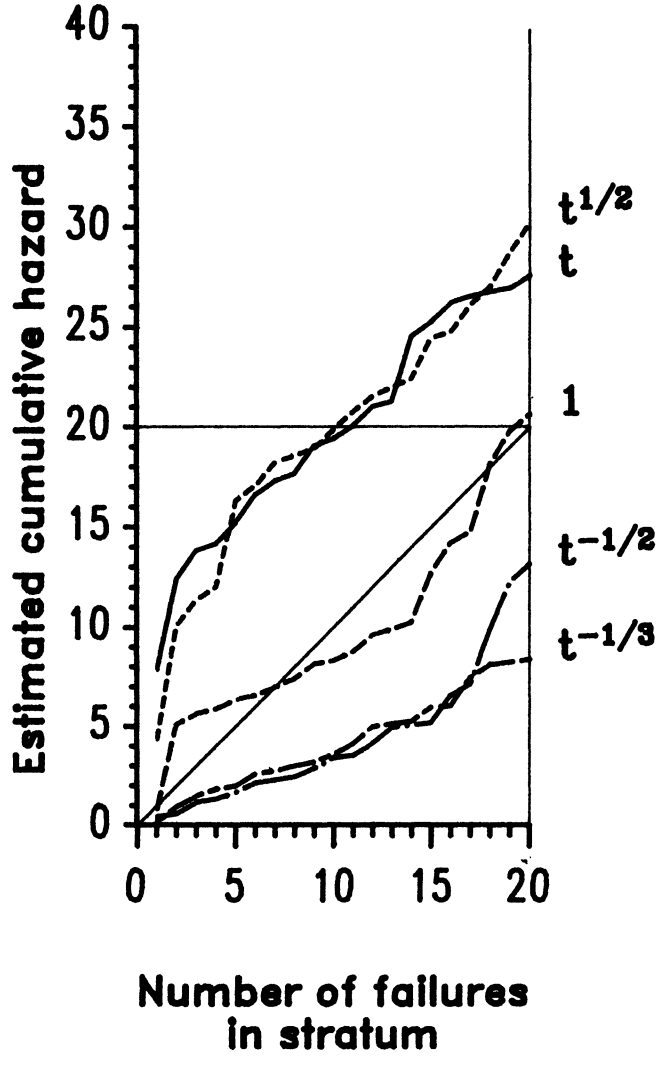

Figure 3. The Behavior of the Diagnostic Curves When It Was Incorrectly Assumed That the Hazards Are Proportional. Five baseline hazards were used: $\lambda_0(t) = t^{p-1}$, $p = 2, \frac{3}{2}, 1, \frac{2}{3}, \frac{1}{2}$. The coefficient estimate was $\hat{\beta} = 1.97$.

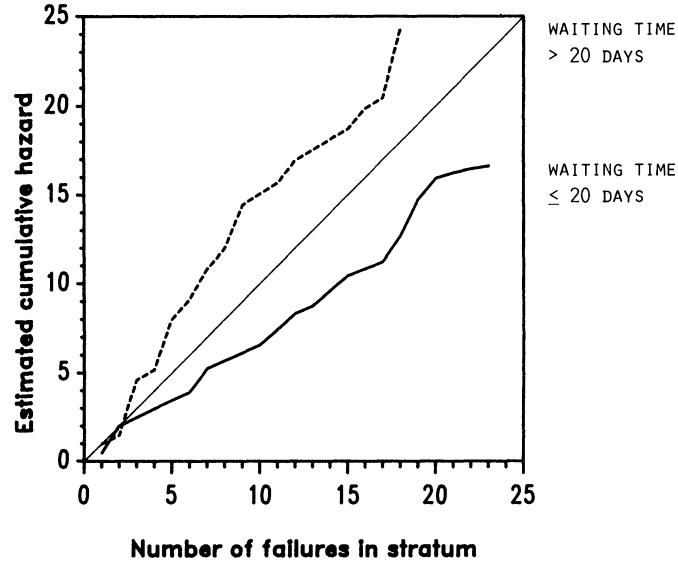

Figure 4. The Diagnostic Curves Describing the Effect of the Waiting Time on the Posttransplant Survival in the Stanford Data (65 patients). The covariates and the corresponding estimated coefficients were AGE (.051), MISMATCH (.470), and SURGERY (−.822).

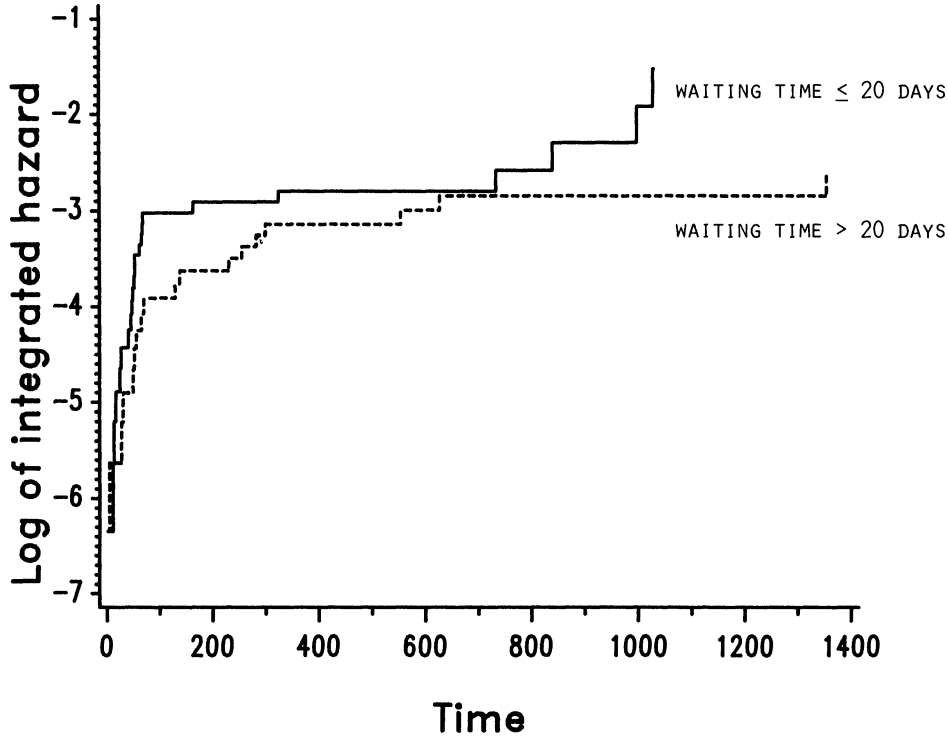

Figure 5. The Logarithmic Cumulative Baseline Hazard Estimates in the Situation Described in Figure 4.

"log $\hat{\Lambda}_s(t)$ versus t" graph of Andersen (1982) (Fig. 5) is also presented. The plots in Figures 4 and 5 are consistent with each other, but Figure 4 seems to deliver the message of nonproportionality more clearly.

As the second example we consider the malignant mel-

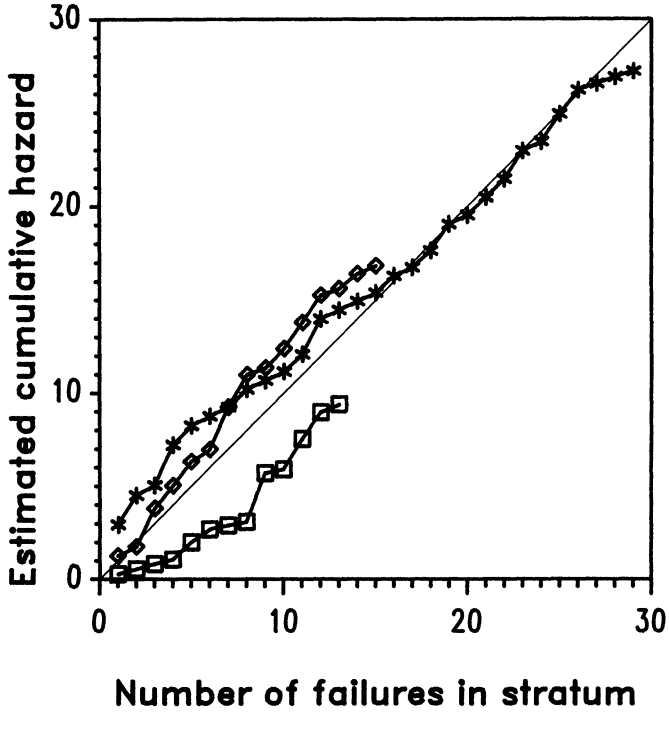

*Figure 6. The Diagnostic Curves Corresponding to the Three Levels of Invasion in the Malignant Melanoma Data (205 patients). For coefficient estimates see Andersen (1982). ◇-◇-◇, invasion level 1; *-*-*, invasion level 2; □-□-□, invasion level 3.*

anoma data of Drzewiecki and Andersen (1982). In particular, we pose the same question as in Andersen (1982): Having fitted an eight-covariate proportional hazards model, what is the role of the level of invasion in surgery (which was not included as a covariate)? Three levels of invasion were considered. These levels also form the natural basis for the stratification in our method. The diagnostic graphs are shown in Figure 6. It is instructive to compare this figure with figure 1 in Andersen (1982). The conclusions are in good agreement with Andersen's.

4. CONCLUDING REMARKS

So far the approach has been almost completely data analytic: apart from the basic martingale property, nothing has been said about the character and magnitude of randomness in the graphs. I now comment on this.

Exact distributional results seem extremely hard to obtain. Two kinds of approximation, however, are suggested fairly easily: (a) considering the standardized differences

$$(N_I(k) - H_I^{\beta_0}(k))/\left(\sum_{i\le k} p_I^{\beta_0}(i)[1 - p_I^{\beta_0}(i)]\right)^{1/2}, \qquad k \ge 1,$$

as (0, 1) normal random variables; (b) considering the spacings

$$H_I^{\beta_0}(\gamma_k^I) - H_I^{\beta_0}(\gamma_{k-1}^I), \qquad k > 1,$$

as unit-exponential and, therefore, their sums $H_I^{\beta_0}(\gamma_k^I)$ as $(k, 1)$ gamma random variables.

The first approximation is motivated by the martingale central limit theorem once we observe that $\Sigma_{i\le k}\, p_I^{\beta_0}(i)[1 - p_I^{\beta_0}(i)]$ $(k \ge 1)$ is the variance process corresponding to the "Bernoulli" counting process N_I. The second approx-

imation is likely to be reasonably accurate if the probabilities $p_I^{\beta_0}(i)$ are small (compare Arjas and Haara, in press); this happens typically when the ratio card(I)/n is small. This theoretical reasoning is complicated by the fact that, as was noted in Section 3, β_0 is unknown in practice and must, therefore, be replaced by the partial maximum likelihood estimate $\hat{\beta}$. Under regularity conditions $\hat{\beta}$ is consistent as $n \to \infty$ (see Andersen and Gill 1982), and so we may, because of continuity, holding I and k fixed, approximate the probabilities $p_I^{\beta_0}(k)$ by $p_I^{\hat{\beta}}(k)$. For a realistic approach to asymptotics, however, one should let card(I) and $K(I)$ grow at the same rate as the sample size n. It follows that a simple continuity argument is not sufficient for establishing how the graphs arising from a correctly specified model behave asymptotically. [The analog to the score-martingale is instructive: When β_0 is replaced by $\hat{\beta}$, the asymptotic limit of the score is no longer a Gaussian martingale but, rather, a transformed Brownian bridge; see Wei (1984) and Haara (1987)].

To get an idea about how marked the differences are between the graphs, based on β_0 with respect to $\hat{\beta}$, a number of graphs with both parameters was drawn. In simulations similar to Figures 1–3, the differences were usually very small. With smaller sample sizes and more parameters they are of course clearer, but even then it seems that the conclusions from the diagnostics would only very rarely depend on whether β_0 or $\hat{\beta}$ is used in the drawing. As an illustration, one set of graphs from a correctly specified model and one from an incorrectly specified model are reproduced here (Fig. 7). For the clarity of the pictures only two groups in each were used.

Although formal goodness-of-fit tests have not been introduced here, it would be of interest to know how accurate is the normal approximation in (a) above, in particular when β_0 is replaced by $\hat{\beta}$, and, in addition, what is the power of the diagnostic method. As a tentative analysis, a number of test trials were run, usually based on 1,000 replicates. These simulations indicate that the normal approximation is actually very good. For an illustration, reproduced here (Fig. 8) are the normal probability plots for the variables

$$D_I = (N_I(\gamma_k^I) - H_I^{\hat{\beta}}(\gamma_k^I))/\left(\sum_{i \le \gamma_k^I} p_I^{\hat{\beta}}(i)[1 - p_I^{\hat{\beta}}(i)]\right)^{1/2},$$

$$I = I_0, \ldots, I_4,$$

choosing the models exactly as in Figures 1 and 2 (n = 100), and k = 20. Again both the correctly specified and incorrectly specified model were tried. For the correctly specified model all five statistics D_I ($I = I_0, \ldots, I_4$) had empirical distributions very close to the (0, 1) normal. When the second covariate was omitted, the distributions were still very close to normal with unit variance, but the means had shifted, separating the distributions clearly. Figure 8 also indicates that the power of the diagnostic curves is quite high in this case. More definite conclusions will require further work.

I close with some remarks concerning the stratification.

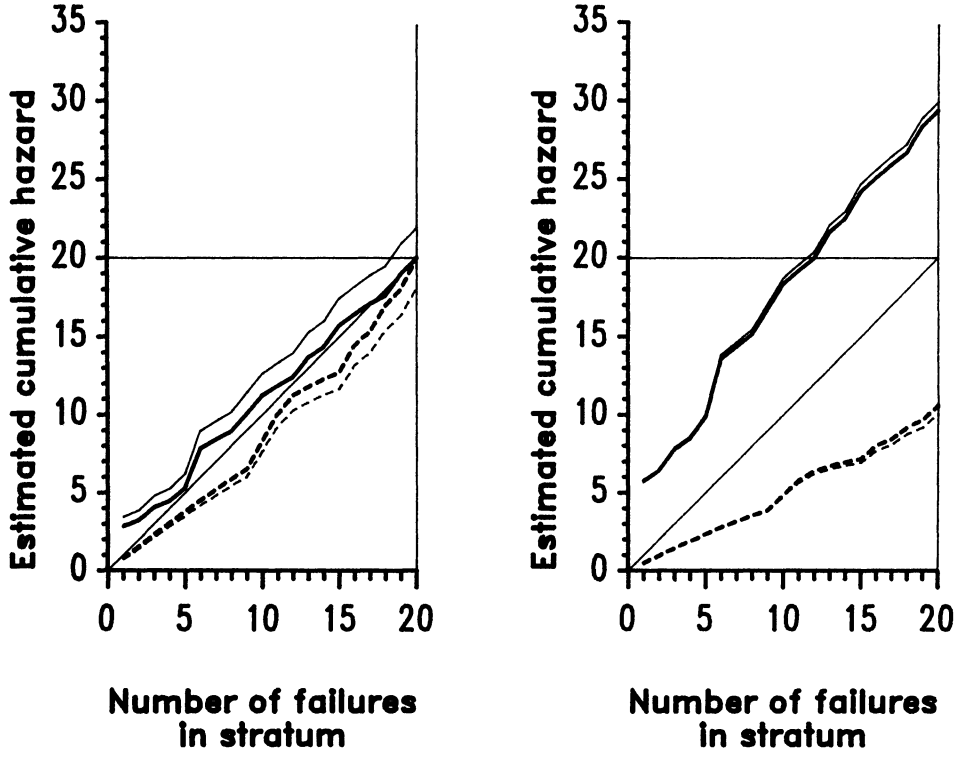

Figure 7. The Difference in the Form of the Diagnostic Curves When the True Simulation Parameter β_0 With Respect to the Estimated Value $\hat{\beta}$ Was Used. There were 40 individuals and two fixed covariates (Z_{j1}, Z_{j2}), where Z_{j1} was evenly spaced over the unit interval in the same way as in Figures 1–3, and Z_{j2} had alternating values 0 and 1. In the incorrect specification of the model the second covariate was omitted. The values of the coefficients were β_0 = (2, 1) (left, fine line), $\hat{\beta}$ = (2.28, 1.27) (left, thick line), β_{01} = 2 (right, fine line), and $\hat{\beta}_1$ = 1.53 (right, thick line).

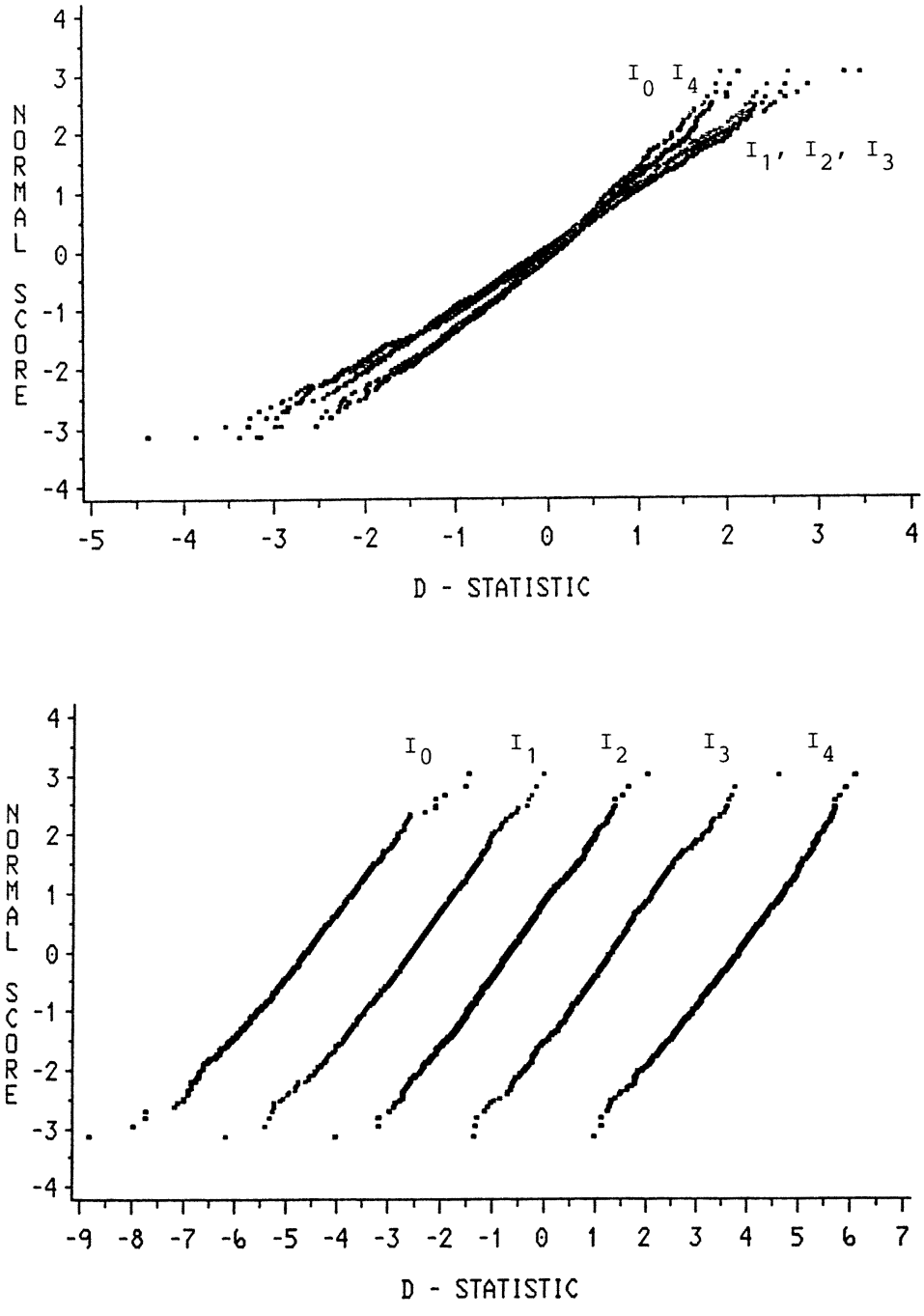

Figure 8. Normal Probability Plots of the D Statistics (defined in Sec. 5). The plots are based on 1,000 replicates from the model used for Figures 1 and 2. The top plot is obtained by fitting the correct model, and in the bottom plot the second covariate has been omitted. Both plots contain the simulated empirical cumulative distribution functions corresponding to the five covariate levels.

As has been pointed out earlier, in practice the stratification will depend on what particular aspect in the model is being considered in the goodness-of-fit problem. (This could be paralleled with the selection of "weights" in defining a test statistic.) Starting from a specific question and the corresponding stratification, the calculation of D_I statistics, in the way explained previously, can give a quantitative idea of the significance of the findings. On the other hand, if one were to explore through many stratifications, a (dependent) sequence of graphs would arise, some probably looking more striking than others. Although such a procedure can sometimes prove useful in indicating possible weaknesses in the model, one should be very cautious in making statements about significance. For such a purpose an omnibus test seems preferable.

[*Received August 1985. Revised June 1987.*]

REFERENCES

Aalen, O. O., and Hoem, J. M. (1978), "Random Time Changes for Multivariate Counting Processes," *Scandinavian Actuarial Journal*, 81–101.

Andersen, P. K. (1982), "Testing Goodness-of-Fit of Cox's Regression and Life Model," *Biometrics*, 38, 67–77.

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for

Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100–1120.
Arjas, E. (1985), Comment on "Counting Process Models for Life History Data: A Review," by P. K. Andersen and Ø. Borgan, *Scandinavian Journal of Statistics*, 12, 150–153.
Arjas, E., and Haara, P. (1987), "A Note on the Asymptotic Normality in the Cox Regression Model," unpublished manuscript.
——— (in press), "A Note on the Exponentiality of Total Hazards Before Failure," *Journal of Multivariate Analysis*.
Barlow, R., and Campo, R. (1975), "Total Time on Test Processes and Applications to the Failure Data Analysis," in *Reliability and Fault Tree Analysis*, eds. S. H. Moolgavkar and R. L. Prentice, Philadelphia: Society for Industrial and Applied Mathematics, pp. 451–481.
Breslow, N. E., Edler, L., and Berger, J. (1984), "A Two-Sample Censored-Data Rank Test for Acceleration," *Biometrics*, 40, 1049–1062.
Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.
——— (1979), "A Note on the Graphical Analysis of Survival Data," *Biometrika*, 66, 188–190.
Crowley, J., and Hu, M. (1977), "Covariance Analysis of Heart Transplant Survival Data," *Journal of the American Statistical Association*, 72, 27–36.
Crowley, J., and Storer, B. E. (1983), Comment on "A Reanalysis of the Stanford Heart Transplant Data," by M. Aitkin, N. Laird, and B. Francis, *Journal of the American Statistical Association*, 78, 277–281.
Drzewiecki, K. T., and Andersen, P. K. (1982), "Survival With Malignant Melanoma. Regression Analysis of Prognostic Factors," *Cancer*, 49, 2414–2419.
Gill, R. D. (1984), "Understanding Cox's Regression Model: A Martingale Approach," *Journal of the American Statistical Association*, 79, 441–447.
Gill, R. D., and Schumacher, M. (1987), "A Simple Test for the Proportional Hazards Assumption," *Biometrika*, 74, 289–300.
Haara, P. (1987), "A Note on the Asymptotic Behaviour of the Empirical Score in Cox's Regression Model for Counting Processes," in *Proceedings of the 1st World Congress of the Bernoulli Society*, Tashkent: VNU Science Press, pp. 139–142.
Hjort, N. L. (1984), "Weak Convergence of Cumulative Intensity Processes When Parameters Are Estimated, With Applications to Goodness-of-Fit Tests in Models With Censoring," research report, Norwegian Computing Center.
Kalbfleisch, J. D. (1974), "Some Extensions and Applications of Cox's Regression and Life Model," presented at the Joint Statistical Meeting in Tallahassee, Florida, March.
Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley.
Kay, R. (1977), "Proportional Hazard Regression Models and the Analysis of Censored Survival Data, *Applied Statistics*, 26, 227–237.
Lagakos, S. W. (1981), "The Graphical Evaluation of Explanatory Variables in Proportional Hazards Regression Models," *Biometrika*, 68, 93–98.
Lagakos, S. W., and Schoenfeld, D. (1984), "Properties of Proportional Hazards Score Tests Under Misspecified Regression Models, *Biometrics*, 40, 1037–1048.
Moreau, T., O'Quigley, J., and Mesbah, M. (1985), "A Global Goodness-of-Fit Statistic for the Proportional Hazards Model, *Applied Statistics*, 34, 212–218.
Schoenfeld, D. (1980), "Chi-Squared Goodness-of-Fit Tests for the Proportional Hazards Regression Model, *Biometrika*, 67, 145–153.
——— (1982), "Partial Residuals for the Proportional Hazards Regression Model," *Biometrika*, 69, 239–241.
Thall, P. F., and Lachin, J. M. (1986), "Assessment of Stratum–Covariate Interactions in Cox's Proportional Hazards Regression Model," *Statistics in Medicine*, 5, 73–84.
Thomas, D. C. (1983), "Non-parametric Estimation and Tests of Fit for Dose–Response Relations," *Biometrics*, 39, 263–268.
Wei, L. J. (1984), "Testing Goodness of Fit for Proportional Hazards Model With Censored Observations," *Journal of the American Statistical Association*, 79, 649–652.