STATISTICS IN MEDICINE, VOL. 14, 1707–1723 (1995)

GRAPHICAL METHODS FOR ASSESSING VIOLATIONS OF THE PROPORTIONAL HAZARDS ASSUMPTION IN COX REGRESSION

KENNETH R. HESS

Department of Biomathematics, Box 237, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, Texas 77030-4095, U.S.A.

SUMMARY

A major assumption of the Cox proportional hazards model is that the effect of a given covariate does not change over time. If this assumption is violated, the simple Cox model is invalid, and more sophisticated analyses are required. This paper describes eight graphical methods for detecting violations of the proportional hazards assumption and demonstrates each on three published datasets with a single binary covariate. I discuss the relative merits of these methods. Smoothed plots of the scaled Schoenfeld residuals are recommended for assessing PH violations because they provide precise usable information about the time dependence of the covariate effects.

INTRODUCTION

The Cox proportional hazards (PH) regression model[1] relates the hazard function for an individual with a vector of covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ to a baseline hazard function via a log-linear function:

$$h(t; \mathbf{x}) = h_0(t) \exp(\mathbf{x}' \boldsymbol{\beta}).$$

The baseline or background hazard function, $h_0(t)$, represents the hazard function for an individual with covariate values all equal to zero. One estimates $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$, a vector of p unknown regression coefficients, without specifying the baseline hazard function, and inference about $\boldsymbol{\beta}$ is possible without making assumptions about the form of the baseline hazard function. In the presence of multiple covariates, the Cox model assumes that the effects of the covariates on the hazard function are multiplicative (that is, additive on the log scale). Thus, the assumption is that joint effect of two covariates is equal to the product of their separate effects. In addition, for an interval-scaled covariate, there is the assumption that the effect of the covariate on the log hazard ratio is linear.

These assumptions are not required, however, for a single binary covariate where the Cox model reduces to

$$h(t; x) = h_0(t) \exp(x\beta).$$

If the variable has values of 0 and 1, the hazard function for group 0 is $h(t; x = 0) = h_0(t)$ (that is, the baseline hazard) and that for group 1 is $h(t; x = 1) = h_1(t) = h_0(t)\exp(\beta)$. One assumes that

CCC 0277–6715/95/151707–17

Received March 1994
Revised September 1994

the ratio of the hazard functions $h_1(t)/h_0(t) = \exp(\beta)$ is constant with respect to time, and thus, the hazard functions $h_0(t)$ and $h_1(t)$ are assumed proportional. Since violations of the PH assumption can lead to incorrect inferences, it is important to check for PH violations.[2]

METHODS

There have been several graphical methods suggested for assessing the proportional hazards assumption:

1. plotting predicted survival curves based on the Cox model along with non-model-based (for example, Kaplan–Meier) estimates of the observed survival curves;
2. plotting the cumulative hazard functions against time and checking for constant ratio;
3. plotting the cumulative hazard functions against each other and checking for constant slope;
4. plotting the log of the cumulative hazard functions against time and checking for parallelism;
5. plotting differences in the log cumulative hazard functions against time and checking for constancy;
6. partitioning the time axis and fitting models separately to each time interval;
7. including time-by-covariate interaction terms in the model and estimating the log hazard ratio function;
8. plotting Schoenfeld's residuals against time to identify patterns.

Although there have been many numerical goodness-of-fit statistics proposed to detect violations of the PH assumption, none of these has gained wide use, and they are limited both by statistical power and the family of alternatives considered.[3–19] In addition, the basis for several of the statistics is the partitioning of the time axis which is equivalent to methods described below. Additional graphical methods for assessing PH appear in Thaler,[20] Mau,[21] Henderson and Milner,[22] and Arjas.[23]

Comparing survival estimates

Recalling that the cumulative hazard function, $H(t)$, relates to the hazard function, $h(t)$, by $H(t) = \int_0^t h(u)\,du$ and that the survival function, $S(t)$, relates to $H(t)$ by $S(t) = \exp(-H(t))$, we can derive the following equivalent expressions for the Cox PH model:

$$h(t) = h_0(t)\exp(\beta x)$$

$$\int_0^t h(u)\,du = \int_0^t h_0(u)\,du \exp(\beta x)$$

$$H(t) = H_0(t)\exp(\beta x)$$

$$\exp[-H(t)] = \exp[-H_0(t)\exp(\beta x)]$$

$$S(t) = \exp(-H_0(t))^{\exp(\beta x)}$$

$$S(t) = S_0(t)^{\exp(\beta x)}.$$

Breslow[24,25] gives an estimate for the cumulative baseline hazard, $H_0(t) = -\log S_0(t)$, based on the Cox PH model

$$\hat{H}_0(t) = \sum_{t_i \leqslant t} \frac{\delta_i}{\sum_{j \in R_i} \exp(\hat{\beta} x_j)},$$

where δ_i is the censoring indicator for the ith individual, and R_i is the set of indices of individuals at risk when the ith individual fails. Additional methods for estimating $H_0(t)$ appear in Kalbfleisch and Prentice[26] and Link.[27] Thus, $\hat{S}_0(t) = \exp(-\hat{H}_0(t))$ gives an estimate of the survival function based on the Cox model. It is possible to assess violations of the PH assumption by comparing survival estimates based on the Cox PH model with estimates computed independently of the model.[1,25,26,29] For a single binary variable, we can compare the Kaplan–Meier[28] estimate for each group with the estimate computed from the Cox model.[1] Clear departures of the two estimates would provide evidence against the PH assumption. This idea extends to multiple binary, polychotomous or interval covariates.[29]

In this approach to comparison of survival estimates, it is not always possible to judge whether discrepancies between 'predicted' (model-based) and 'observed' (non-model-based) survival estimates result from sampling fluctuations or from real trends.[25,29] The presentation of confidence intervals may help. Computation of confidence intervals for the Kaplan–Meier estimates is not difficult,[30] but intervals for the estimates from the Cox model involve fairly complex calculations.[27] Even if calculation of the intervals were trivial, however, presentation of both curves and both sets of intervals leads to a cluttered graph. It is possible, however, to present the confidence intervals on separate plots.

Plotting cumulative hazard functions

The PH assumption implies that $S(t) = S_0(t)^{\exp(\beta x)}$; thus, the survival curves are powers of one another. We can use this observation as a rudimentary check of the PH assumption through inspection of the Kaplan–Meier survival curve estimates, since the crossing of survival curves indicates departure from the assumption. The PH assumption also implies that $H(t) = H_0(t)\exp(\beta x)$, and, thus, that the cumulative hazard curves have a constant ratio. Here again, crossing curves indicate violations of the PH assumption. Since $H(t) = -\log S(t)$, we can use the $-\log$ transformation of the Kaplan–Meier estimate for this assessment. The PH assumption further implies that $\log H(t) = \log H_0(t) + \beta x$; thus, we can rewrite the PH model as

$$\log[-\log S(t)] = \log[-\log S_0(t)] + \beta x.$$

Therefore, under PH, plots of $\log[-\log \hat{S}_i(t)]$ (or equivalently, plots of $\log \hat{H}_i(t)$) are roughly parallel.[5,25,26] For a single binary covariate, it is possible simply to take $\log[-\log]$ transformations of the Kaplan–Meier estimates and check for equidistance between the curves.

For the two-sample case, several authors[4,9,14,31,32] suggest plotting $H_1(t)$ versus $H_0(t)$. Under PH, $H_1(t) = \theta H_0(t)$ where $\theta = \exp(\beta)$ is constant over t. Thus, the recommended plot (sometimes called an H–H plot) yields a straight line with slope θ and zero intercept. Muenz[33] recommends plotting $\hat{\omega}(t)$ versus t, where

$$\hat{\omega}(t) = \hat{H}_1(t)/\hat{H}_0(t).$$

Under PH, $\omega(t) = \theta$. Schumacher[34] suggests plotting $\hat{\gamma}(t)$ versus t, where

$$\hat{\gamma}(t) = \log[-\log \hat{S}_1(t)] - \log[-\log \hat{S}_0(t)].$$

Since $H(t) = -\log S(t)$, $\hat{\gamma}(t) = \log \hat{H}_1(t) - \log \hat{H}_0(t) = \log(\hat{H}_1(t)/\hat{H}_0(t)) = \log \hat{\omega}(t)$. Under PH, $\gamma(t) = \log \theta = \beta$, a constant with respect to t. Dabrowska *et al.*[35] recommend plotting $\hat{\phi}(t)$ against t, where

$$\hat{\phi}(t) = (\hat{H}_1(t) - \hat{H}_0(t))/\hat{H}_0(t).$$

Since $(\hat{H}_1(t) - \hat{H}_0(t))/\hat{H}_0(t) = \hat{H}_1(t)/\hat{H}_0(t) - 1$, $\hat{\phi}(t) = \hat{\omega}(t) - 1$. Under PH, $\phi(t) = \theta - 1$. The three functions $\omega(t)$, $\gamma(t) = \log \omega(t)$, and $\phi(t) = \omega(t) - 1$ are closely related and yield similar plots. Given the widespread use of the log[− log] survival plots and the fact that $\gamma(t) = \log[-\log S_1(t)] - \log[-\log S_0(t)]$, the plot of $\hat{\gamma}(t)$ versus t seems a reasonable choice. This plot gives a direct assessment of the PH assumption. Rather than a comparative assessment of parallelism between two curves, we have the assessment of constancy of a single curve . Under PH, this plot is constant over t, centred around $\hat{\beta}$, the estimated log hazard ratio. In general, however, if $\beta(t) = \log(h_1(t)/h_0(t))$, $\gamma(t) \neq \beta(t)$ since $\log(H_1(t)/H_0(t)) \neq \log(h_1(t)/h_0(t))$. Furthermore, since $\gamma(t)$ is the logarithm of the ratio of the integrals of $h_1(t)$ and $h_0(t)$, $\gamma(t)$ and $\beta(t)$ have a complex relationship. Hence, we cannot infer the shape of $\beta(t)$ from the shape of $\gamma(t)$.

Partitioning the time axis and fitting models to each interval

A straightforward alternative to the simple PH model is to partition the time axis and fit separate PH models to each time interval.[5,25,29,36] Anderson and Senthilselvan,[37] and O'Quigley and Pessione[38] suggest a simple model with just two intervals, but others have pursued more general approaches.[3,4,39] Gore *et al.*[40] describe a process for assessing the time dependence of the influence of clinical covariates by fitting simple PH models in distinct time intervals and comparing the coefficients (that is, log hazard ratios) across intervals. For each interval, one deletes patients whose deaths or censoring occur prior to the interval, and one censors the survival times for patients who live through the interval at the end of the interval.[10,25] Several authors have observed that this process is equivalent to defining time-dependent covariates, which are indicator variables for the time intervals.[15,29,36,41-43] Some authors have also noted that one can implement the indicator variables with use of existing computer programs for Cox PH regression. Harrell and Lee[29] describe the process as 'stratification on time', a term also used by Kalbfleisch and Prentice.[36]

For k time intervals, consider the introduction of $k - 1$ time-dependent indicator variables to partition the time axis into intervals: $(\tau_0, \tau_1), \ldots, (\tau_{k-1}, \tau_k)$ with $\tau_0 = 0$ and $\tau_k = \infty$.[10,41] Define the indicator variables such that $I_j = 1$ for $\tau_{j-1} \leqslant t < \tau_j$ and $I_j = 0$ otherwise. The Cox model extends to

$$h(t, x) = h_0(t) \exp\{(\beta + \zeta_j)x\} \text{ for } \tau_{j-1} \leqslant t < \tau_j.$$

This model allows the effect of the covariate x to vary as a step function and to take on the value $\beta + \zeta_j$ in the jth interval.[10] To create four intervals, we can use three values of the follow-up times (T_1, T_2, T_3) to define three new indicator variables (I_2, I_3, I_4) such that $I_j = 1$ for $T_{j-1} \leqslant t < T_j$ and $I_j = 0$ otherwise. For a single binary covariate, this yields a new Cox model with a log hazard ratio function (LHRF) = $\log(h_1(t)/h_0(t))$ where

$$\text{LHRF} = \beta_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 I_4.$$

The corresponding LHRF estimates are β_1 for the first interval, $\beta_1 + \beta_2$ for the second, $\beta_1 + \beta_3$, for the third, and $\beta_1 + \beta_4$ for the fourth. Given the covariance between the regression parameters,

it is straightforward to compute approximate confidence intervals for these estimates. Furthermore, a test that $\beta_2 = \beta_3 = \beta_4 = 0$ is a test of the PH assumption against the specified step-function alternative.

While several authors discuss the problem of choosing the number and location of the breakpoints between time intervals, they make few recommendations. The most common recommendations are to have roughly comparable numbers of events in each interval and to avoid having too few events in any given interval.[15,25] This first recommendation is equivalent to choosing breakpoints that are quantiles of the death times. This means, however, that the results depend on the censoring mechanism, which may be undesirable. An alternative is to select breakpoints that are quantiles of the observation times (that is, including both death and censoring times), which still leaves the problem of selection of the quantiles. The number and location of the breakpoints should depend on the amount of data and the detail needed about the time dependence. Although some authors suggest *post hoc* selection of the breakpoints after initial inspection of the data, this has the undesirable effect of vitiating any statistical inference about the time dependence.

Modelling time-by-covariate interactions

A time-by-covariate interaction occurs when the effect of an explanatory variable on survival changes with time. We can use Cox models with terms to model such interactions to check the PH assumption.[1,5,25,44] We can model a time-by-covariate interaction for a covariate, x, as a product term in the Cox regression equation

$$h(t;x) = h_0(t)\exp(\beta_1 x + \beta_2 x f(t)),$$

where $f(t)$ is a function of time. The corresponding LHRF for individuals differing in x is

$$\text{LHRF} = \Delta(x)(\beta_1 + \beta_2 f(t)) = \beta(t),$$

where $\Delta(x) = (x_1 - x_2)$ is the difference in the x values. For binary x, $\text{LHRF} = \beta_1 + \beta_2 f(t)$. Selecting the correct functional form for $f(t)$ seems a critical factor in this approach. Linear ($f(t) = a + bt$), logarithmic ($f(t) = a + b\log(t)$), and exponential ($f(t) = a + b\exp(t)$) are commonly used monotonic functions of time.[1,25,37,44] One uses the constants a and b primarily for computational convenience. Modelling non-monotonic time-dependence has received little attention in the literature. In one notable exception, Stablein *et al.*,[45] fit a quadratic polynomial function of time ($at + bt^2$). As noted in the previous section, we may approximate $f(t)$ by a step function that is constant over specified time intervals. Spline functions of time have recently been described by Gray[46] and Hess.[47] These latter methods avoid the problem of having to specify a functional form for $f(t)$, but create the problem of having to specify the number and locations of the knots of the spline functions.

Whether we approximate $f(t)$ by a parametric function, a step function, or spline function, we can plot the estimated relative hazard function and corresponding confidence intervals against time to determine the magnitude of the violation of the PH assumption. Furthermore, a test that $\beta_2 = 0$ is a test of the PH assumption against the alternative specified by $f(t)$. In moderate-to-small data sets, there may be insufficient information to judge the precise nature of $f(t)$. One approach is to plot both the step function estimates and one or more simple parametric estimates to gauge the general form of $f(t)$.

Plotting Schoenfeld residuals

Schoenfeld[48] defined partial residuals for the Cox model that do not depend on time so that one can plot the ith residual against t_i to detect violations of the PH assumption.[5] Let R_i denote the set of indices of those under observation when the ith individual fails (that is, the risk set at t_i). As we drop patients from the risk set, the mean covariate value $E(x_i)$ of those remaining (that is, $E(x_i \mid R_i)$) changes. Under PH, the Cox model yields

$$E(x_i \mid R_i) = \frac{\sum_{k \in R_i} x_k \exp(x_k \beta)}{\sum_{k \in R_i} \exp(x_k \beta)}.$$

Let $\hat{\beta}$ denote the maximum partial likelihood estimate under PH, and let $\hat{E}(x_i \mid R_i)$ denote $E(x_i \mid R_i)$ with $\hat{\beta}$ substituted for β. Now define the partial residual at t_i as

$$\hat{r}_i = x_i - \hat{E}(x_i \mid R_i).$$

If the PH assumption holds, $E(\hat{r}_i) \approx 0$ and a plot of $\hat{r}_i$ versus t_i centres around 0.[48] Trends in this plot reflect time dependence in the covariate effects and, hence, violations of the PH assumption. For a dichotomous covariate coded 0 or 1, the residuals are $1 - E(x_i \mid R_i)$ for $x_i = 1$ and $0 - E(x_i \mid R_i)$ for $x_i = 0$, yielding two horizontal bands of residuals.[48] The number of points in each band is equal to the number of unique death times in the corresponding group.

Grambsch and Therneau[49] describe a scale adjustment for Schoenfeld's residuals. The scaled version of the residuals permits the interpretation of the smoothed residuals as a non-parametric estimate of the LHRF. Use of the approximate adjustment they refer to as average variance standardization yields a scaled Schoenfeld residual,

$$\hat{r}_i^* = \hat{\beta} + \hat{r}_i d \widehat{\text{var}}\ (\hat{\beta}),$$

where $\hat{\beta}$ is the maximum partial likelihood estimate under PH, $\hat{r}_i$ is the Schoenfeld residual (also under PH), $\widehat{\text{var}}\ (\hat{\beta})$ is the estimated variance of $\hat{\beta}$, and d is the total number of events where individuals from both groups remain at risk. If the PH assumption holds, a plot of $\hat{r}_i^*$ versus t_i centres around $\hat{\beta}$.

Trends in the scatterplots of the Schoenfeld residuals are often difficult to ascertain, especially with binary covariates where there are only two horizontal bands of residuals present. Superposition of the results of a smoothing procedure can dramatically improve the interpretability of the residual plots.[50] Smoothing helps describe the pattern of dependence of the mean of a response variable y (in this case the residuals) as a function of a variable x (in this case time). The only assumption is that the mean of y at each value of x, $E(y \mid x)$, varies as a smooth function of x. LOWESS (locally-weighted scatterplot smoothing) employs iterated weighted least squares with a robustness feature that, after an initial locally-weighted smoothing, identifies and down-weights outliers in successive smoothings.[51] A characteristic quantity associated with LOWESS is the local span, window or bandwidth, defined as the fraction (f) of points in the neighbourhood around x_i, a given value of x, used in the estimation of $E(y \mid x_i)$. As the fraction increases, the smoothness increases, and local fluctuations decrease. A trade-off exists between increased bias from too much smoothing and increased variance from too little smoothing. A detailed overview of smoothing appears in the recent book by Hastie and Tibshirani.[52] As discussed by Grambsch and Therneau,[49] it is possible to compute confidence intervals for the smoothed residuals. These intervals help to separate real PH violations from random fluctuations in the data.

Table I. Ovarian carcinoma dataset: time to progression of disease in days

Low-grade tumours ($N = 15$): 28, 89, 175, 195, 309, 377$^+$, 393$^+$, 421$^+$, 447$^+$, 462, 709$^+$, 744$^+$, 770$^+$, 1106$^+$, 1206$^+$
High-grade tumours ($N = 20$): 34, 88, 137, 199, 280, 291, 299$^+$, 300$^+$, 309, 351, 358, 369, 369, 370, 375, 382, 429$^+$, 451, 1119$^+$

$^+$ indicates censored observation

RESULTS

To compare the graphical methods of detecting violations of the PH assumption, I use three previously published and analysed datasets. Each of these datasets contains a single binary covariate, and since the only assumption required to model such data with a Cox regession model is the PH assumption, they focus attention on this assumption. Each dataset demonstrates a different departure from PH; the first has survival curves that diverge toward the end of follow-up, the second has survival curves that intersect toward the end of follow-up, and the third has survival curves that cross near the middle of follow-up. I apply the various methods for graphical detection of violations of the PH assumption to each of these datasets in turn, and I compare the information provided by the various methods.

Ovarian cancer dataset

Fleming *et al.*[53] present data for 35 patients with limited stage II or IIIA ovarian cancer (Table I). The patients are divided into two groups based on grade of disease. Fifteen patients had low-grade or well-differentiated cancer, and 20 had high-grade or undifferentiated cancer. Nine (60 per cent) of the patients with low-grade tumours had censored data, while four (20 per cent) with high-grade tumours had censored data. Fleming *et al.*,[53] Breslow *et al.*,[8] and Gill and Schumacher[14] each found that these data violated the PH assumption.

For the simple Cox model, I introduced a single binary covariate coded 0 for low-grade tumours and 1 for high-grade tumours. The estimated log hazard ratio ($\hat{\beta}$) is 1·119 (standard error: 0·497). Survival (that is, freedom from progression) estimates based on this Cox model agree with the Kaplan–Meier estimates before 100 days and after 350 days, but not in the middle phase (Figures 1(*a*) and 1(*b*)). The cumulative hazard functions (Figure 1(*c*)) are similar for the first 300 days, but thereafter, the curve for the high-grade tumour patients increases dramatically whereas that for the low-grade tumour patients remains fairly constant. This sharp divergence is also apparent in the *H–H* plot (Figure 1(*d*)).

The divergence of the cumulative hazard functions is apparent when plotted on the log scale (Figure 2(*a*)). The plot of the log of the ratio of the cumulative hazard functions (Figure 2(*b*)) shows the increase directly. Superimposing a LOWESS smooth (bandwidth = 1/2) helps clarify the trend. To partition the time axis, I computed quartiles (165·5, 309, and 371·2 days) and terciles (199 and 369 days) of death time, and I introduced time-dependent indicator variables into the Cox model. Estimation for the quartiles would not converge. Inspection of the data indicated a lack of low-grade tumour patients in the third interval. Changing the second cutpoint from 309 to 308 and thereby moving a low-grade tumour patient from the second interval into the third interval corrected the problem. The resulting log hazard ratio estimates are 0·100, 0·058, 1·843 and 2·410 for the quartiles and − 0·322, 1·863 and 2·546 for the terciles. Introduction of a linear

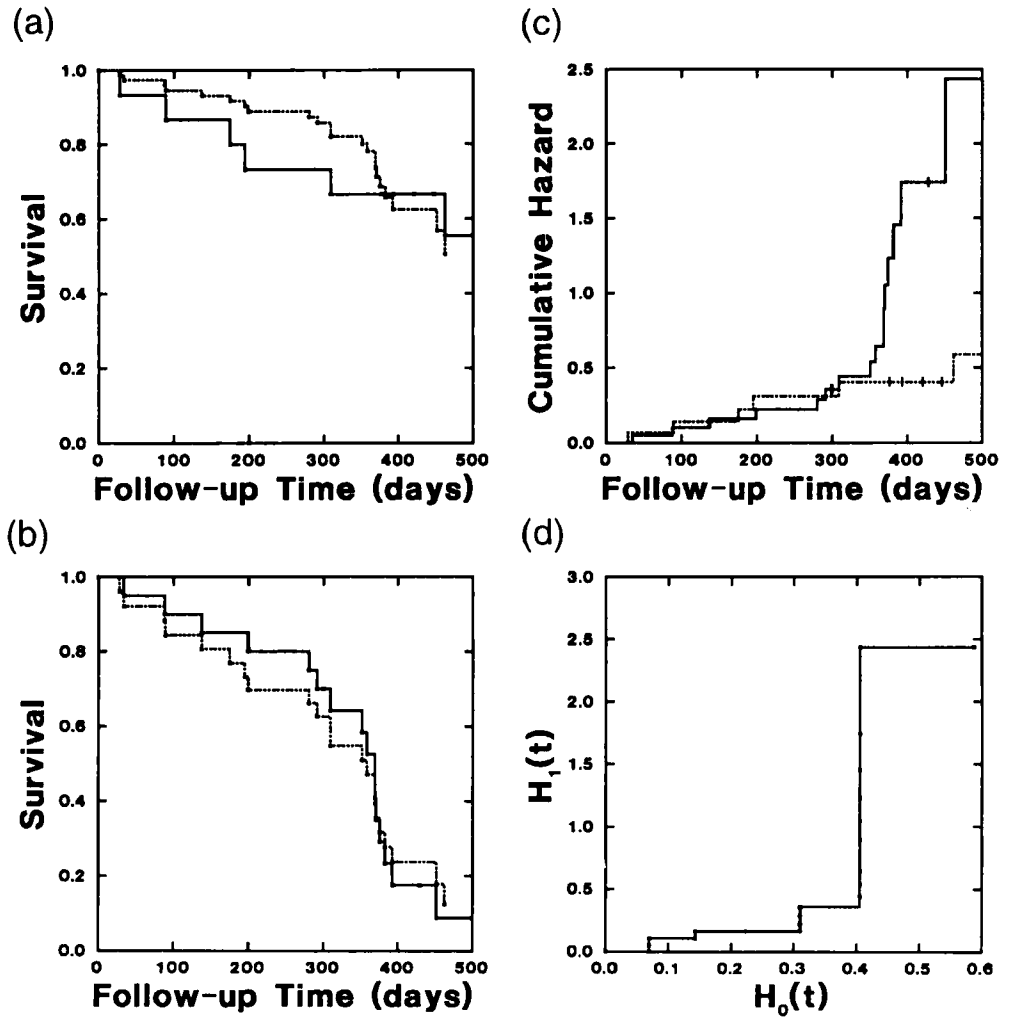

Figure 1. Ovarian cancer data: (*a*) Low-grade tumour patients, Cox model estimates (dotted line) and Kaplan–Meier estimates (solid line); (*b*) High-grade tumour patients, Cox model estimates (dotted line) and Kaplan–Meier estimates (solid line); (*c*) Cumulative hazard functions, low-grade (dotted line), high-grade (solid line), and (*d*) *H–H* plot

time-covariate interaction yields $\hat{\beta}_1 = -1{\cdot}022$ and $\hat{\beta}_2 = 0{\cdot}008$. A quadratic time-covariate interaction gives $\hat{\beta}_1 = -0{\cdot}349$ and $\hat{\beta}_2 = 0{\cdot}0000164$. A plot of the corresponding LHRF estimates appears in Figure 2(*c*). The LOWESS-smoothed (bandwidth = 2/3) scaled Schoenfeld residual plot shows a generally increasing trend starting negative and crossing zero at about 250 days (Figure 2(*d*)).

Gastric cancer dataset

Stablein *et al.*[45] present data for 90 patients with locally advanced, non-resectable gastric cancer (Table II). Half were treated with chemotherapy alone ($N = 45$) and half with chemotherapy plus radiation ($N = 45$). There were 8 (18 per cent) censored in each group. This dataset was

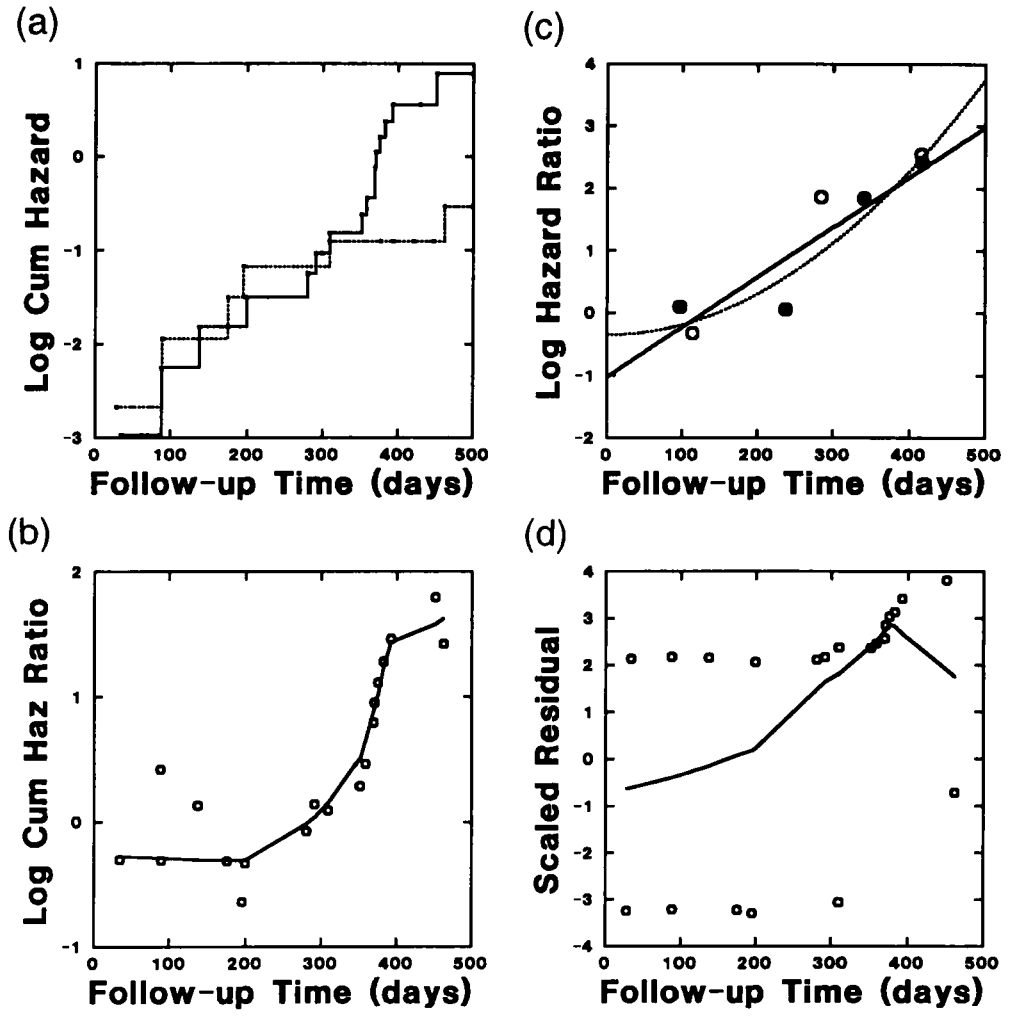

Figure 2. Ovarian cancer data: (*a*) Log cumulative hazard functions, low-grade (dotted line) and high-grade (solid line); (*b*) Smoothed plot of the log of the ratio of the cumulative hazard functions; (*c*) Cox model LHRF estimates from Cox models with time covariate interactions, terciles step function (open circles), quartile step function (solid circles), linear (solid line) and quadratic (dotted line), and (*d*) Smoothed plot of the scaled Schoenfeld residual

Table II. Gastric carcinoma dataset: survival time in days

Chemotherapy and radiation ($N = 45$): 17, 42, 44, 48, 60, 72, 74, 95, 103, 108, 122, 144, 167, 170, 183, 185, 193, 195, 197, 208, 234, 235, 254, 307, 315, 401, 445, 464, 484, 528, 542, 567, 577, 580, 795, 855, 882$^+$, 892$^+$, 1031$^+$, 1033$^+$, 1306$^+$, 1335$^+$, 1366, 1452$^+$, 1472$^+$

Chemotherapy ($N = 45$): 1, 63, 105, 129, 182, 216, 250, 262, 301, 301, 342, 354, 356, 358, 380, 381$^+$, 383, 383, 388, 394, 408, 460, 489, 499, 524, 529$^+$, 535, 562, 675, 676, 748, 748, 778, 786, 797, 945$^+$, 955, 968, 1180$^+$, 1245, 1271, 1277$^+$, 1397$^+$, 1512$^+$, 1519$^+$

$^+$ indicates censored observation

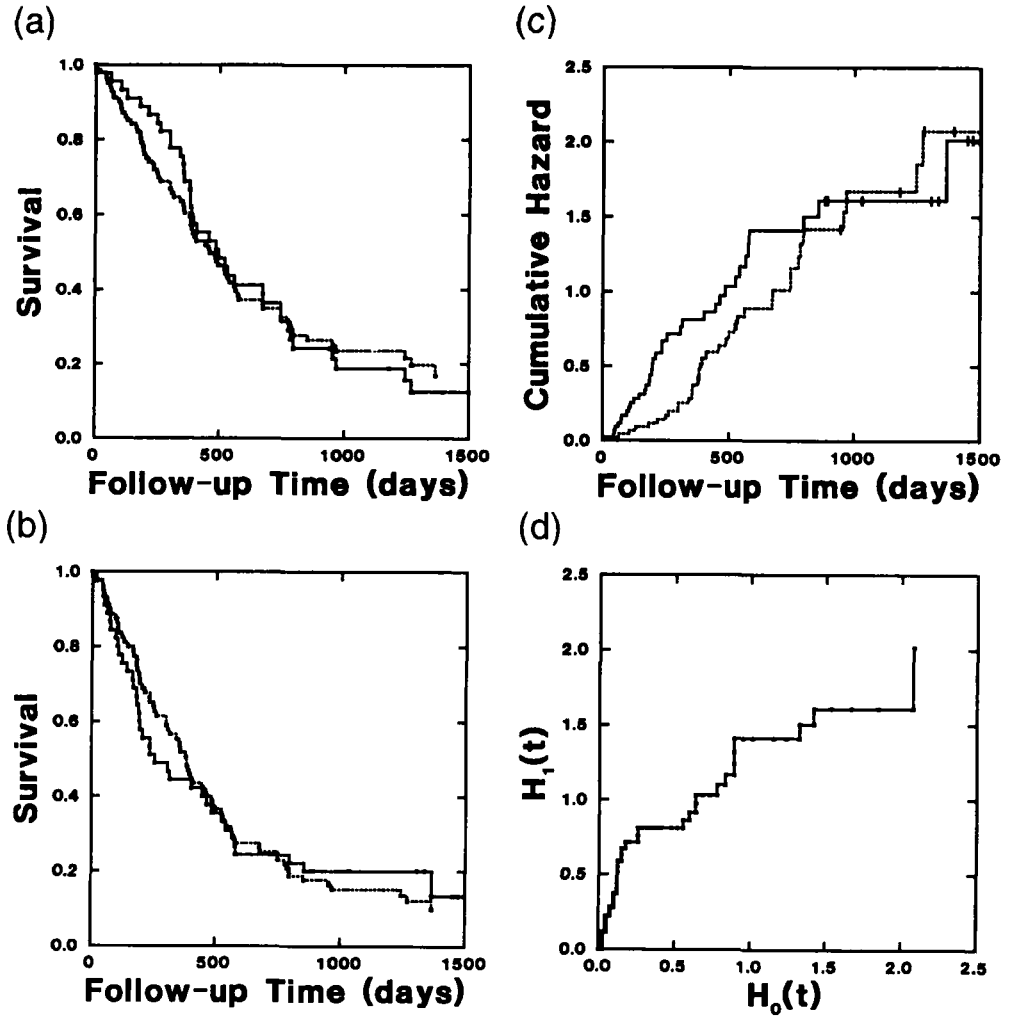

Figure 3. Gastric cancer data: (*a*) Chemotherapy alone, Cox model estimates (dotted line) and Kaplan–Meier estimates (solid line); (*b*) Chemotherapy plus radiation, Cox model estimates (dotted line) and Kaplan–Meier estimates (solid line); (*c*) Cumulative hazard functions, chemotherapy alone (dotted line) and chemotherapy plus radiation (solid line), and (*d*) *H–H* plot

subsequently analysed by Stablein and Koutrouvelis,[11] Mantel and Stablein[54] and Moreau *et al.*,[10] all of whom verified the assessment of Stablein *et al.*[45] of crossing hazard functions. In addition, O'Quigley and Pessione[15] showed significant departure from the PH model.

For the Cox model, I introduced a single binary covariate coded 0 for chemotherapy alone and 1 for chemotherapy plus radiation. The estimated log hazard ratio ($\hat{\beta}$) is 0·267 (standard error 0·233). Survival estimates based on this Cox model agree with the Kaplan–Meier estimates between 400 and 900 days but not in the early and late phases (Figures 3(*a*) and (*b*)). The cumulative hazard functions (Figure 3(*c*)) are well separated until about 700 days when they converge. This convergence is evident in the *H–H* plot (Figure 3(*d*)) in a flattening of the slope.

The convergence of the cumulative hazard functions is also apparent when plotted on the log scale (Figure 4(*a*)). The LOWESS-smoothed plot of the log of the ratio of the cumulative hazard

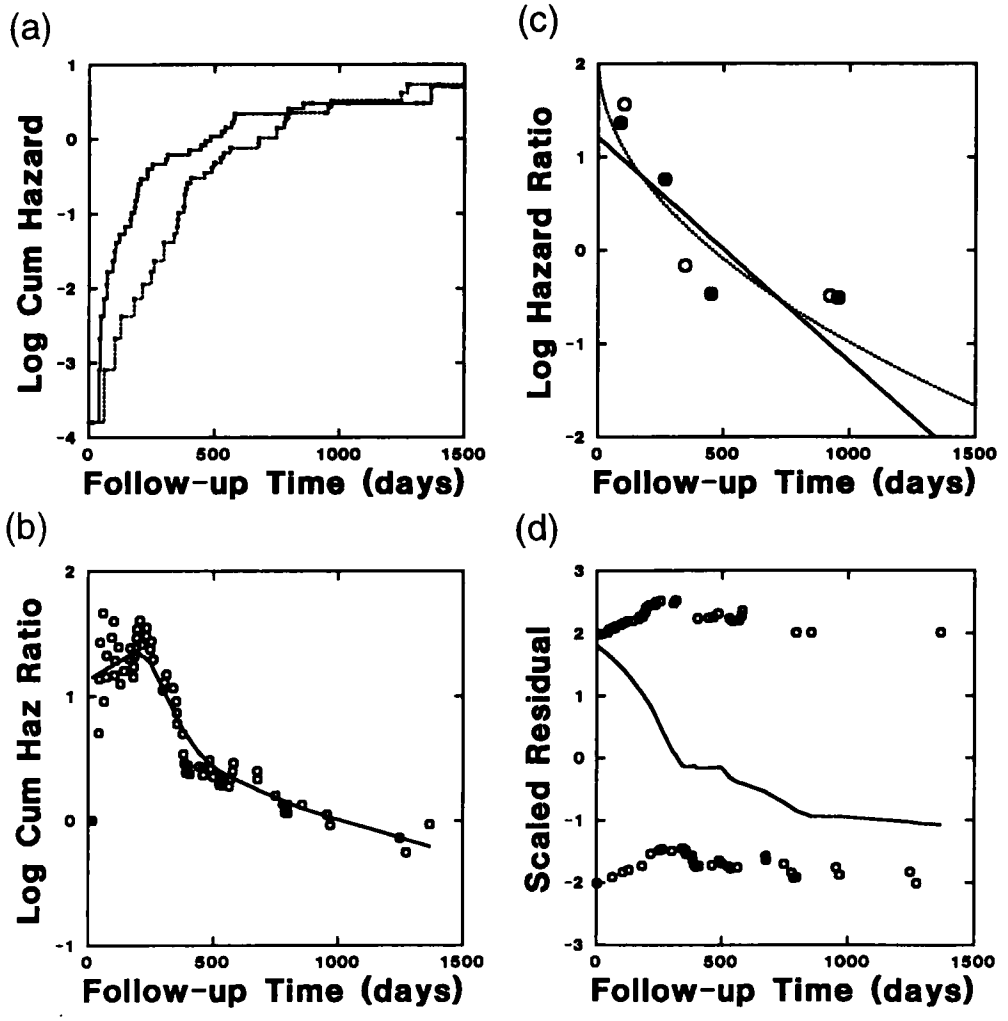

Figure 4. Gastric cancer data: (*a*) Log cumulative hazard functions, chemotherapy alone (dotted line) and chemotherapy plus radiation (solid line); (*b*) Smoothed plot of the log ratio of the cumulative hazard functions; (*c*) Cox model LHRF estimates from Cox models with time covariate interactions, tercile step function (open circles), quartile step function (solid circles), linear (solid line) and square root (dotted line), and (*d*) Smoothed plot of the scaled Schoenfeld residual

functions (Figure 4(*b*)) shows the convergence through a sharp decline. I partitioned the time axis on death-time quartiles (179, 355 and 547 days) and terciles (208 and 484 days) with use of time-dependent indicator variables. The log hazard ratio estimates are 1·368, 0·762, − 0·465 and − 0·507 for the quartiles and 1·565, − 0·163, and − 0·487 for the terciles. A linear time-covariate interaction gives $\hat{\beta}_1 = 1{\cdot}209$ and $\hat{\beta}_2 = -0{\cdot}002$. A square root function for time yields $\hat{\beta}_1 = 2{\cdot}059$ and $\hat{\beta}_2 = -0{\cdot}096$. Figure 4(*c*) shows a plot of the time–covariate interactions. The LOWESS-smoothed scaled Schoenfeld residual plot shows a generally decreasing trend, starting positive, crossing zero just before 300 days, and then demonstrating an abrupt decline in the rate of decrease (Figure 4(*d*)).

Table III. Bile duct cancer dataset: survival time in days

Treated ($N = 22$):	30, 67, 79$^+$, 82$^+$, 95, 148, 170, 171, 176, 193, 200, 221, 243, 261, 262, 263, 399, 414, 446, 446$^+$, 464, 777
Control ($N = 25$):	57, 58, 74, 79, 89, 98, 101, 104, 110, 118, 125, 132, 154, 159, 188, 203, 257, 257, 431, 461, 497, 723, 747, 1313, 2636

$^+$ indicates censored observations

Bile duct cancer dataset

Fleming *et al.*[53] present data for 47 patients with bile duct cancer (Table III). These patients were followed to determine whether those treated with a combination of radiation treatment and chemotherapy ($N = 22$) survive longer than a control group ($N = 25$). Three (14 per cent) of the treated patients had censored data, while none (0 per cent) of the control patients did. Fleming *et al.*,[53] Breslow *et al.*,[8] and O'Quigley and Pessione[37] noted the crossing survival curves and the effect of this phenomenon on the usual two sample tests for comparing survival curves (that is, the Cox–Mantel logrank test and the Gehan–Breslow generalized Wilcoxon test). They each proposed tests more powerful than the usual tests in the absence of PH, but they did not specifically assess the PH violations of these data.

For the Cox PH model, I introduced a binary covariate coded 0 for control and 1 for treated patients. The estimated log hazard ratio ($\hat{\beta}$) is $-0{\cdot}07$ with standard error 0·32. The survival estimates based on this Cox model agree with the Kaplan–Meier estimates between 250 and 450 days but not before or after (Figures 5(*a*) and 5(*b*)). The cumulative hazard functions (Figure 5(*c*)) cross at about 400 days. The crossing is not readily apparent from the *H–H* plot (Figure 5(*d*)) although its sigmoid shape snakes around the unity line.

The crossing cumulative hazard functions are evident when plotted on the log scale (Figure 6(*a*)). The LOWESS-smoothed plot of the log of the ratio of the cumulative hazard functions (Figure 6(*b*)) gives direct evidence of a divergence-convergence-crossing pattern. I partitioned the time axis on death-time quartiles (105·5, 190·5 and 410·2 days) and terciles (132 and 261 days) with use of time-dependent indicator variables. The resulting log hazard ratio estimates are $-0{\cdot}858$, $-0{\cdot}764$, 0·886 and 0·667 for the quartiles and $-1{\cdot}325$, 0·210 and 1·035 for the terciles. A linear time-covariate interaction yields $\hat{\beta}_1 = -0{\cdot}750$ and $\hat{\beta}_2 = 0{\cdot}003$, and a logarithmic time-covariate interaction yields $\hat{\beta}_1 = -4{\cdot}219$ and $\hat{\beta}_2 = 0{\cdot}796$. Figure 6(*c*) shows a plot of the time–covariate interactions. The LOWESS-smoothed scaled Schoenfeld residual plot (Figure 6(*d*)) starts as negative but increases rapidly, crossing zero at about 200 days. The increase stops abruptly at about 250 days, followed by a general decline toward zero.

DISCUSSION

Although others have reviewed graphical methods for assessing PH violations,[8,25,26,29,36,44] there has been no systematic comparison of the available methods. These methods have advantages and disadvantages and succeed to varying degrees in assisting the data analyst in detection of violations of the PH assumption. The apparent differences between model-based and non-model-based survival estimates are not particularly large for the three datasets analysed. On closer inspection, however, each plot demonstrates differences greater than 10 per cent over some segments of the follow-up. Although inspection of the non-model-based survival estimates can show evidence for violations of the PH assumption, it is generally difficult to assess visually

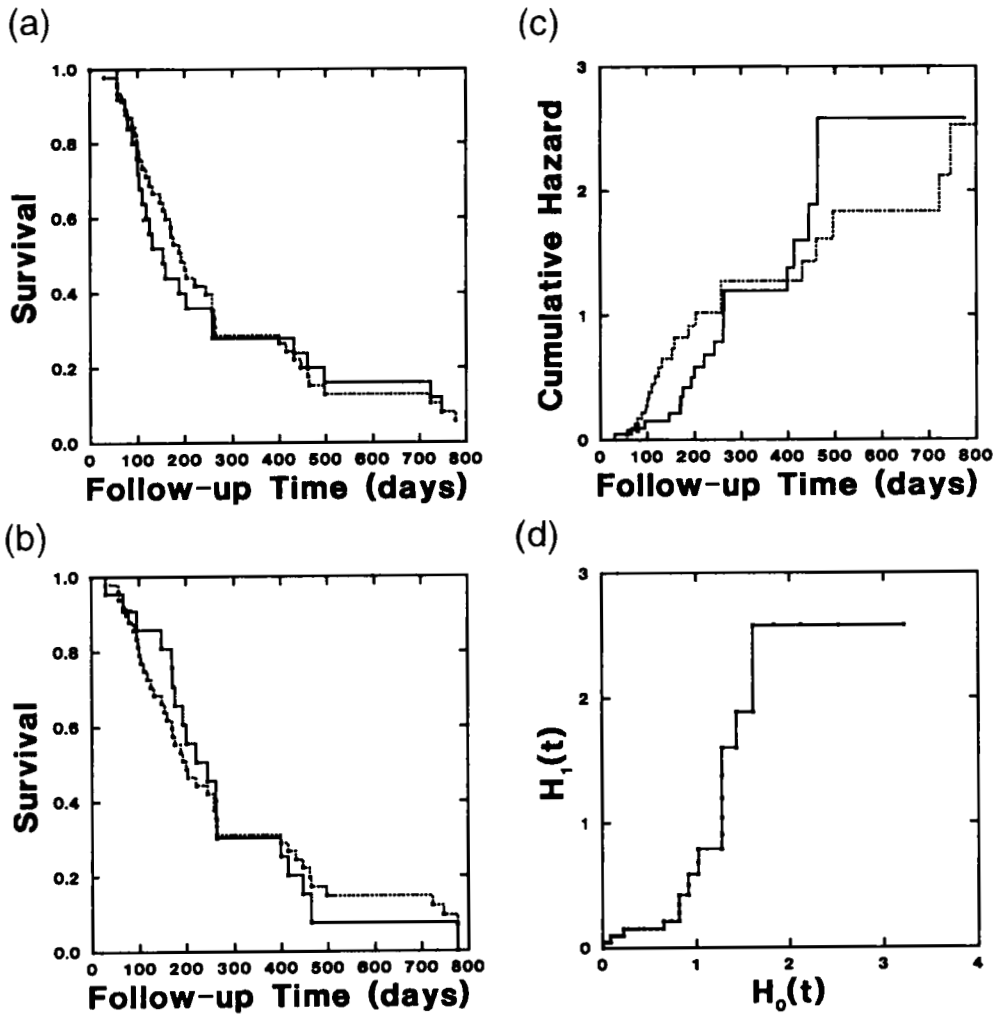

Figure 5. Bile duct cancer data: (*a*) Control patients, Cox model estimates (dotted line) and Kaplan–Meier estimates (solid line); (*b*) Treated patients, Cox model estimates (dotted line) and Kaplan–Meier estimates (solid line); (*c*) Cumulative hazard functions, control (dotted line) and treated (solid line), and (*d*) H–H plot

whether curves are powers of one another. It is similarly difficult to assess if plots of the cumulative hazard function have a constant ratio. The H–H plot in which one plots the estimated cumulative hazard function of one group against the other, requires visual assessment of whether a curve has a constant slope, which is not a trivial task. Plots of $\log[-\log S(t)]$ or, equivalently, $\log H(t)$ versus t, require assessment of whether two curves are parallel. While this is arguably more reliable than assessment of whether or not two curves have constant ratios or power, it can be difficult. Plotting the point-by-point differences against t makes the task somewhat easier. Smoothing these plots makes trends more apparent. These trends, however, do not give direct information about the shape of log hazard ratio function since they represent the logarithm of the ratio of the cumulative hazard functions and not the logarithm of the ratio of the hazard functions.

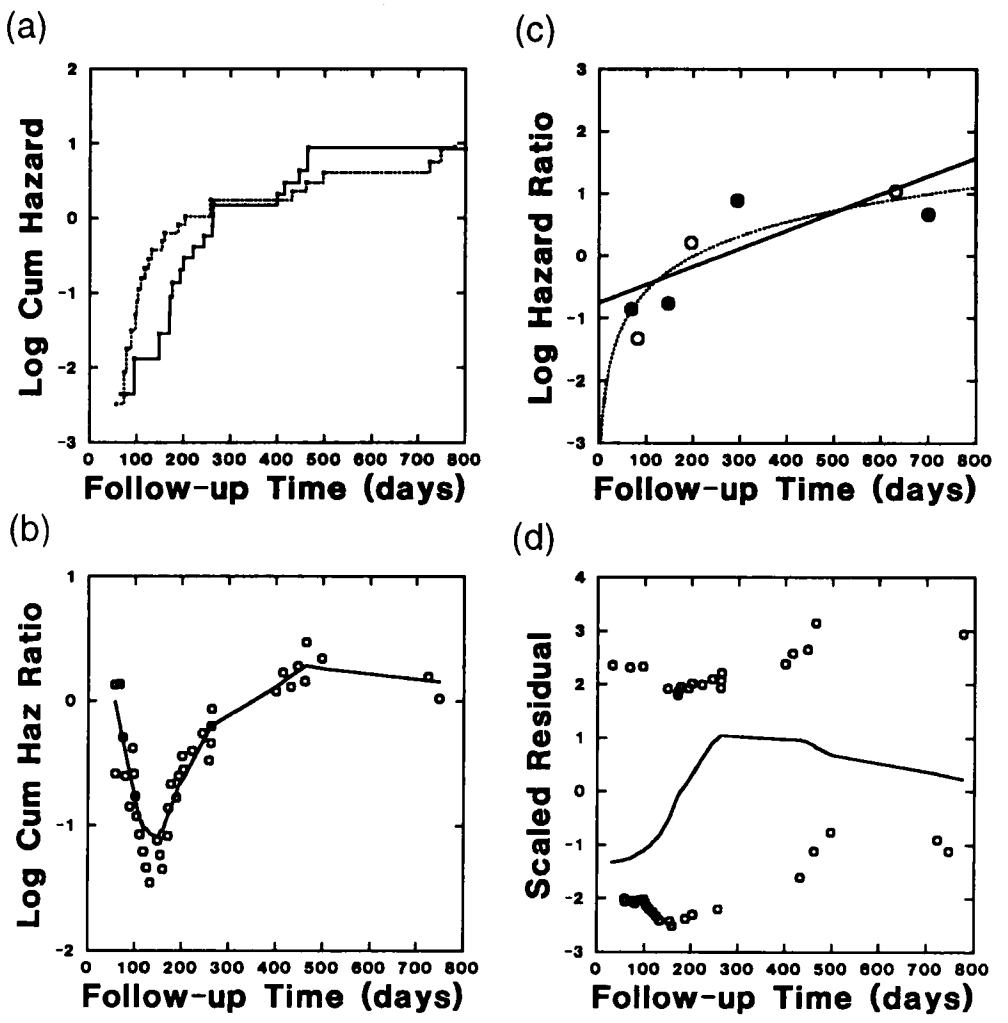

Figure 6. Bile duct cancer data: (*a*) Log cumulative hazard functions, control (dotted line) and treated (solid line); (*b*) Smoothed plot of the log ratio of the cumulative hazard functions; (*c*) Cox model LHRF estimates form Cox models with time covariate interactions, tercile step function (open circles), quartile step function (solid circles), linear (solid line) and logarithm (dotted line), and (*d*) Smoothed plot of the scaled Schoenfeld residual

The use of time-dependent indicator variables to partition the time axis seems useful in general and for non-proportional hazards in particular. With sufficient data, we can use them to detect non-linearities in the time dependencies. Fitting separate models to each time interval is also useful for datasets with many covariates.[40] Although this approach assumes common break-points for all covariates, it provides a useful first approximation for general time-dependence. The possible dependence of the results on the number and location of breakpoints between time intervals suggests that one should exercise care in their choice. Use of specific functions of time to model time-by-covariate interactions is problematic because of the necessity of specifying the function. One requires some prior knowledge or expectation of the form of the time dependence. In practice, investigation is often limited to simple monotonic functions. The plotting of several simple step function and parametric LHRF estimates gives a general impression of the shape of

the time dependence, even with insufficient data available to determine the precise function of time.

The smoothed Schoenfeld residual plots seem useful. They provide relatively precise information about the time dependence of the covariate effects. This should prove useful for selection of the functional form for a time-by-covariate interaction or for selection of breakpoints in the partitioning of the time axis. The Grambsch and Therneau[49] scaled version of the Schoenfeld residual seems even more useful than the original. This residual has the same scale as the LHRF; thus, we can interpret the smoothed plots as non-parametric estimates of the LHRF. We can investigate the effect of bandwidth choice on the estimate by using several different bandwidths. Hastie and Tibshirani[55] pursue the more general concept of time-varying coefficients models with use of a penalized partial likelihood method. Exploration of methods of assessing PH violations based on the score process and martingale residuals appear in Lin *et al.*[56] and Therneau *et al.*[57]

All eight of these methods for graphical assessment of violations of the PH assumptions in two-sample datasets proved helpful to some degree. The various plots of the survival and cumulative hazard estimates aided in detection of departures from the PH assumptions but provide no method for dealing with the violation in the two-sample case. Furthermore, with violations detected, these methods provide little information about how to modify the Cox model to account for the violation. In the presence of multiple covariates, one can use a stratified Cox model to overcome the PH assumption for a particular covariate. The various methods for fitting time-by-covariate interactions in a Cox model are useful both for detection and treatment of PH violations. Trends evident from smoothed Schoenfeld residual plots are helpful for selecting the function of time used in the time-by-covariate interaction. If we cannot adequately summarize the trend with a simple parametric form, then we could employ spline functions.[47]

Although I presented only datasets with single binary covariates in this paper, one can also use these methods for datasets with interval-scaled, polychotomous and multiple covariates. These more general settings, however, introduce topics such as checking the linearity and main-effects assumptions, topics beyond the scope of this paper. Furthermore, although I used real datasets to demonstrate the proposed methods in this report, it would be helpful to explore simulated data with known hazard ratio functions to confirm that the methods recover these functions.

ACKNOWLEDGEMENTS

The author thanks Robert J. Hardy, Barry R. Davis and Gary F. Marks for their advice and encouragement. The author also thanks Susan Steagall and Laura Leeah for their secretarial assistance. This work is based in part on the author's doctoral dissertation at The University of Texas School of Public Health.

REFERENCES

1. Cox, D. R. 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society*, **34**, 187–220 (1972).
2. Lagakos, S. W. and Schoenfeld, D. A. 'Properties of proportional-hazards score tests under misspecified regression models', *Biometrics*, **40**, 1037–1948 (1984).
3. Schoenfeld, D. 'Chi-squared goodness-of-fit tests for the proportional hazards regression model', *Biometrika*, **67**, 145–153 (1980)
4. Andersen, P. K. 'Testing goodness of fit of Cox's regression and life model', *Biometrics*, **38**, 67–77 (1982).
5. Kay, R. 'Goodness of fit methods for the proportional hazards regression model: A review', *Revue d'Épidémiologie et Santé Publique*, **32**, 185–198 (1984).
6. Nagelkerke, N. J. D., Oosting, J. and Hart, A. A. M. 'A simple test for goodness of fit of Cox's proportional hazards model', *Biometrics*, **40**, 483–486 (1984).

7. Schumacher, M. and Vaeth, M. 'On a goodness-of-fit test for the proportional hazards model', *EDV in Medizin und Biologie*, **15**, 19–23 (1984).
8. Breslow, N. E. Edler, L. and Berger, J. 'A two-sample censored-data rank test for acceleration', *Biometrics*, **40**, 1049–1062 (1984).
9. Wei, L. J. 'Testing goodness of fit for proportional hazards model with censored observations', *Journal of the American Statistical Association*, **79**, 649–652 (1984).
10. Moreau, T., O'Quigley, J. and Mesbah, M. 'A global goodness-of-fit statistic for the proportional hazards model', *Applied Statistics*, **34**, 212–218 (1985).
11. Stablein, D. M. and Koutrouvelis, I. A. 'A two-sample test sensitive to crossing hazards in uncensored and singly censored data', *Biometrics*, **41**, 643–652 (1985).
12. Moreau, T., O'Quigley, J. and Lellouch, J. 'On D. Schoenfeld's approach for testing the proportional hazards assumption', *Biometrika*, **73**, 513–515 (1986).
13. Harrell, F. E. 'The PHGLM procedure' *in SUGI Supplemental Library User's Guide, Version 5*, SAS Institute, Inc., Cary, NC, 1986.
14. Gill, R. D. and Schumacher, M. 'A simple test of the proportional hazards assumption', *Biometrika*, **74**, 289–300 (1987).
15. O'Quigley, J. and Pessione, F. 'Score tests for homogeneity of regression effect in the proportional hazards model', *Biometrics*, **45**, 135–144 (1989).
16. Lin, D. Y. 'Goodness-of-fit analysis for the Cox regression model based on a class proportional hazards model', *Journal of the American Statistical Association*, **86**, 725–728 (1991).
17. Horowitz, J. L. and Neumann, G. R. 'A generalized moments specification test of the proportional hazards model', *Journal of the American Statistical Association*, **87**, 234–240 (1992).
18. Le, C. T. and Zelterman, D. 'Goodness of fit tests for proportional hazards regressions models', *Biometrical Journal*, **5**, 557–566 (1992).
19. Gray, R. J. 'Spline based tests in survival analysis', *Biometrics*, **50**, 640–652 (1994).
20. Thaler, H. T. 'Nonparametric estimation of the hazard ratio', *Journal of the American Statistical Association*, **79**, 290–293 (1984).
21. Mau, J. 'On a graphical method for the detection of time-dependent effects of covariates in survival data', *Applied Statistics*, **35**, 245–255 (1986).
22. Henderson, R. and Milner, A. 'Aalen plots under proportional hazards', *Applied Statistics*, **40**, 401–409 (1991).
23. Arjas, E. 'A graphical method for assessing goodness of fit in Cox's proportional hazards mdoel', *Journal of the American Statistical Association*, **83**, 204–212 (1988).
24. Breslow, N. Discussion on 'Regression models and life tables' by D. R. Cox, *Journal of the Royal Statistical Society, Series B*, **34**, 216–217 (1972).
25. Breslow, N. 'Methods for identifying mortality risk factors in longitudinal studies', *in* Vallin, J., Pollard, H. J. and Heligman, L. (eds.), *Methodologies for the Collection and Analysis of Mortality Data*, Ordina Editions, Belgium, 1984, pp. 367–391.
26. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980, pp. 86–89.
27. Link, C. L. 'Confidence intervals for the survival function using Cox's proportional-hazard model with covariates', *Biometrics*, **40**, 601–610 (1984).
28. Kaplan, E. L. and Meier, P. 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association*, **53**, 457–481 (1958).
29. Harrell, F. E. and Lee, K. L. 'Verifying assumptions of the Cox proportional hazards model', *SUGI II: Proceedings of the Eleventh Annual SAS Users Group International Conference*, 823–828 (1986).
30. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980, p. 15.
31. Fisher, N. I. 'Graphical methods in nonparametric statistics: a review and annotated bibliography', *International Statistical Review*, **51**, 25–58 (1983).
32. Schumacher, M. 'Evaluation of nonproportional treatment effects in cancer clinical trials', *Cancer Investigation*, **8**, 91–98 (1990).
33. Muenz, L. R. 'Comparing survival distributions: a review for nonstatisticians. II', *Cancer Investigation*, **1**, 537–545 (1983).
34. Schumacher, M. 'Two-sample tests of Cramér-von Mises- and Kolmogorov-Smirnov-type for randomly censored data', *International Statistical Review*, **52**, 263–281 (1984).

35. Dabrowska, D. M., Doksum, K. A. and Song, J-K. 'Graphical comparison of cumulative hazards for two populations', *Biometrika*, **76**, 763–773 (1989).
36. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980, p. 214.
37. Anderson, J. A. and Senthilselvan, A. 'A two-step regression model for hazard functions', *Applied Statistics*, **31**, 44–51 (1982).
38. O'Quigley, J. and Pessione, F. 'The problem of a covariate-time qualitative interaction in a survival study', *Biometrics*, **47**, 101–115 (1991).
39. O'Quigley, J. and Moreau T. 'Testing the proportional hazards regression model against some general alternatives', *Revue d' Épidémiologie et Santé Publique*, **3–4**, 199–205 (1984).
40. Gore, S. M. Pocock, S. J. and Kerr, G. R. 'Regression models and non-proportional hazards in the analysis of breast cancer survival', *Applied Statistics*, **33**, 176–195 (1984).
41. Simon, R. 'Use of regression models: statistical aspects' *in* Buyse, M. E., Staquet, M. J. and Sylvester, R. J. (eds.), *Cancer Clinical Trials – Methods and Practice*, Oxford University Press, Oxford, 1984, pp. 444–466.
42. Gray, R. J. 'Some diagnostic methods for Cox regression models through hazard smoothing', *Biometrics*, **46**, 93–102 (1990).
43. Liang, K-Y., Self, S. and Liu, X. 'The Cox proportional hazards model with change point: An epidemiologic application', *Biometrics*, **46**, 783–793 (1990).
44. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980, pp. 134–135.
45. Stablein, D. M., Carter, W. H. and Novak, J. W. 'Analysis of survival data with nonproportional hazard functions', *Controlled Clinical Trials*, **2**, 149–159 (1981).
46. Gray, R. J. 'Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis', *Journal of the American Statistical Association*, **87**, 942–951 (1992).
47. Hess, K. R. 'Assessing time-by-covariate interactions in proportional hazards regression model using cubic spline functions', *Statistics in Medicine*, **13**, 1045–1062 (1994).
48. Schoenfeld, D. 'Partial residuals for the proportional hazards regression model', *Biometrika*, **69**, 239–241 (1982).
49. Grambsch, P. M. and Therneau, T. M. 'Proportional hazard tests and diagnostics based on weighted residuals', *Biometrika*, **81**, 515–526 (1994).
50. Pettitt, A. N. and bin Daud, I. 'Investigating time dependence in Cox's proportional hazards model', *Applied Statistics*, **39**, 313–329 (1990).
51. Cleveland, W. S. 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association*, **74**, 829–836 (1979).
52. Hastie, T. and Tibshirani, R. *Generalized Additive Models*, Chapman and Hall, New York, 1990, pp. 9–81.
53. Fleming, T. R., O'Fallon, J. R., O'Brien, P. C. and Harrington, D. P. 'Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data', *Biometrics*, **36**, 607–625 (1980).
54. Mantel, N. and Stablein, D. M. 'The crossing hazard function problem', *The Statistician*, **37**, 59–64 (1988).
55. Hastie, T. and Tibshirani, R. 'Varying-coefficient models', *Journal of the Royal Statistical Society, Series B*, **55**, 757–796 (1993).
56. Lin, D. Y., Wei, L. J. and Yin, Z. 'Checking the Cox model with cumulative sums of martingale-based residuals', *Biometrika*, **80**, 557–572 (1993).
57. Therneau, T. M., Grambsch, P. M. and Fleming, T. R. 'Martingale-based residuals for survival models', *Biometrika*, **77**, 147–160 (1990).