Journal of the Korean
Statistical Society
Vol. 23, No. 2, 1994

Goodness of Fit Tests of Cox's Proportional Hazards Model[†]

HaeHiang Song[1] and SunHo Lee[2]

ABSTRACT

Graphical and numerical methods for checking the assumption of proportional hazards of Cox model for censored survival data are discussed. The strengths and weaknessess of several goodness of fit tests for the proportional hazards for the two-sample problem are evaluated with Monte Carlo simulations, and the tests of Schoenfeld (1980), Andersen (1982), Wei (1984), and Gill and Schumacher (1987) are considered. The goodness of fit methods are illustrated with the survival data of patients who had chronic liver disease and had been treated with the endoscopy injection sclerotherapy. Two other examples of data known to have nonproportional hazards are also used in the illustration.

KEYWORDS: Goodness of fit tests, Proportional hazards, Cox model, Survival data.

[†]The first author's research was supported by a grant from CMC Research Funds and the second author's research was supported by a grant from Korea Science and Engineering Foundation.

[1]Department of Biostatistics, Catholic University Medical College, Seoul, 137-701, Korea.

[2]Department of Mathematics, Sejong University, Seoul, 133-747, Korea.

1. INTRODUCTION

The use of proportional hazards regression of Cox (1972) in the medical literature has grown considerably in the last two decades. The Cox model is particularly useful in examining treatment comparisons based on the time to some event of interest while adjusting for the effects of concommitant variables. The so-called prognostic variables that are significantly associated with the survival are identified with the Cox model, and the analysis methodology of Cox can handle the censored data which are frequent in medical researches.

Checking of model assumptions is an important aspect of Cox regression analyses, but surprisingly little attention has been paid to this problem; few articles reporting on the application of Cox model on survival data actually perform goodness of fit tests of assumptions. It is relatively easy to detect from diagnostic graphs very serious deviations in the proportional hazards assumption. But it is generally difficult to determine, with graphical methods alone, whether some discrepancies from the proportionality indicated by the plots are so important that the proportional hazards Cox model must be rejected. In such situations, concrete test statistics are needed to support the graphical investigations. Several numerical methods have been suggested in the last decade. This paper reviews and discusses graphical and numerical methods that have been proposed for assessing the Cox model assumption.

Survival data of patients with chronic liver disease and two other examples of medical data are described in section 2. Exploratory analyses with survival and hazard functions are done in section 3. Methods for checking Cox model assumption are set in section 4. Application to medical examples and the results from a Monte Carlo study are described in section 5, and further discussions are given in section 6.

2. MEDICAL DATA

The strengths and weaknesses of various goodness of fit test statistics are best illustrated by examples of data. Therefore, three sets of medical data are

used as examples among which two sets are already known in the literature to have nonproportionality of hazards by diagnostic graphs or by a certain goodness of fit test.

The first data set consisted of 133 patients with chronic liver disease who were treated at the Saint Mary's Hospital, Daejun, Korea, from July, 1986 to August, 1992. Since variceal bleeding remained a major cause of death among those patients, obliteration of varicose would seem to be a logical preventive measure to decrease morbidity and mortality, and thus the endoscopic injection sclerotherapy (EIS) was initiated for this purpose on them. Of the 133 patients aged from 19 to 64, 75 had the EIS by a planned regimen (regular EIS group), and 58 had the EIS whenever variceal hemorrhage occurred (episodic EIS group). We reviewed data retrospectively in order to determine whether there was any beneficial effect of the sclerotherapy by a planned regimen compared to the episodic management of the bleeding. We also wanted to identify important prognostic factors, or covariates, that affected survival.

Patients had pretreatment liver function tests that included measurements of total serum bilirubin (BILI), globulin (GLOB), serum albumin (ALB), α-fetoprotein (AFP), prothrombin time (PT) and direct serum bilirubin (DIR). AFP was recoded as 0 if a value was less than 20mg/dl and as 1 otherwise. Log transformed value of BILI(mg/dl) was used. The severity of the liver disease (PSCORE) was graded using a numerical scoring system modified from Pugh and co-workers (1973), and it was a summarization of patient's status of encephalopathy, ascites, BILI, ALB and PT. We also included age (AGE) for its predictive significance.

The second data set consisted of patients with gastric carcinoma treated either by chemotherapy alone or chemotherapy plus radiation. The data set was used by Carter et al. (1983) as an example of nonproportional hazards.

The third data set was used by Wei (1984), which was a part of data given in Hoel (1972). Two groups of male mice were given 300 rads of radiation and placed, respectively, in a germ-free and in a normal environment. Effects of two different environments were compared on developing thymic lymphoma.

3. EXPLORATORY ANALYSES WITH SURVIVAL AND HAZARD FUNCTIONS

Let T represent the random survival time with survival function

$$S(t) = \exp\Big[-\int_0^t \lambda(u)du\Big],. \tag{3.1}$$

where $\lambda(t)$ is the hazard function of time t. By the definition of $S(t) = \Pr(T > t)$, equation (3.1) is equivalent to the definition of the hazard function,

$$\lambda(t) = \lim_{h\to 0} h^{-1} \Pr\Big(t \le T < t + h \Big| T > t\Big). \tag{3.2}$$

Thus, $\lambda(t)$ is the instantaneous rate of death immediately after t, given survival until t, and conveys a precise information concerning the intensity with which deaths occur through time. For this reason, the hazard function plays an important role in the survival modeling.

Survival times may be censored for some patients whose death has not yet occurred when the study is terminated or who are lost during the study. It will be assumed throughout that death and censoring are determined by an independent mechanism. The diagnostic plots of survival and hazard functions are basic tools in survival analysis. The method of Kaplan-Meier (1958) can be used to estimate the survival function, $S(t)$, for the censored data. Figure 1 shows the Kaplan-Meier survival functions for the two treatment groups of chronic liver desease data. The log rank test (Peto and Peto, 1972) for the equality of two survival curves is nonsignificant ($\chi^2 = 2.563$, d.f. $= 1$, $p = 0.11$). Thus, no beneficial effect of regular EIS group over episodic EIS group is recognized.

Consider now the problem of estimating hazard functions. An examination of hazard functions before testing equality of survival is sometimes important, because the log rank test is locally most powerful against proportional hazards alternatives. Although the test statistic may still be used when the proportional hazards assumption does not hold, it will not be very sensitive in detecting survival differences when they exist, particularly if the hazard functions

cross (Pepe and Fleming, 1989). The estimation procedure, however, makes the hazard function difficult to plot directly. An easiest way is to assume that the hazard function is constant over fixed intervals of 4 months, say. (For a shorter interval, the estimated hazard functions become too variable to decipher its pattern.) If d_i is the total number of deaths and B_i is the total time spent by patients in the ith interval, then the estimated hazard function is d_i/B_i (Kalbfleisch and Prentice, 1980, p.16). The estimated hazard functions are plotted for the two treatment groups of chronic liver disease data in Figure 2, and they show a fluctuating pattern.

Nonproportional hazards or unusual hazard patterns are sometimes revealed by hazard plots (Anderson and Senthilselvan, 1982; Gore et al., 1984), but these two studies are of large data sets. Normally the estimated hazard functions from small to moderate data sets are invariably fluctuating or unstable, and thus one cannot obtain a useful information concerning the shapes or the proportionality of hazard functions. For the chronic liver disease data, the estimated hazard functions of Figure 2 do not convey any information.

4. ASSESSING THE ASSUMPTION OF PROPORTIONAL HAZARDS

In this section we discuss graphical and numerical methods of examining the validity of the assumptions of proportional hazard functions. Two commonly used graphical techniques will be considered. As for numerical techniques for checking the adequacy of Cox proportional hazards model, Schoenfeld (1980), Andersen (1982), Wei (1984), and Gill and Schumacher (1987) have proposed. There is also a method proposed by Moreau, O'Quigley and Mesbah (1985). However, the statistic suggested by Moreau et al. is the same as Schoenfeld's (1980) for the two-sample problem, and hence this method is not explored further.

4.1. Graphical methods.

The regression model of Cox (1972) assumes the hazard function for a patient with covariate vector $z = (z_1, \ldots, z_k)'$ to be $\lambda(t; z) = \lambda_0(t) \exp\left(\beta' z\right)$, where $\lambda_0(t)$, an unspecified function of time, is in fact the underlying hazard at $z = 0$, and β is a vector of k unknown regression coefficients. Cox model implies the proportional hazards for two patients with covariate vectors z_1 and z_2, and assumes that covariates have a multiplicative effect on the hazard function.

Suppose that $k+1$ covariates are considered and the proportional hazards assumption holds for the first k covariates $z = (z_1, \ldots, z_k)'$, but the proportional hazards assumption is tested for z_{k+1}. For an indicator variable z_{k+1}, the proportional hazards assumption implies that the hazard function may be formulated as

$$\lambda(t; z) = \lambda_0(t) \exp\left(\beta_{k+1} z_{k+1}\right) \exp\left(\beta' z\right).$$

And, for a quantitative variable z_{k+1}, we first divide the possible values of z_{k+1} into p strata and accordingly the patients, depending on their values of z_{k+1}. By using the strata model of Kalbfleisch and Prentice (1980, Chapter 4), the hazard function for a patient in sth stratum is given by

$$\lambda_s(t; z) = \lambda_{0s}(t) \exp\left(\beta' z\right) \quad s = 1, \ldots, p,$$

where the underlying hazard functions, $\lambda_{01}(t), \cdots, \lambda_{0p}(t)$, are allowed to be arbitrary and are completely unrelated. Then, when the assumption of proportional hazards holds for the variable z_{k+1}, plotting the logarithm of the cumulative underlying hazard in each group (referred also as stratum),

$$\log \Lambda_s(t) = \log \int_0^t \lambda_{0s}(u) du + \beta' z, \quad s = 1, \ldots, p,$$

versus t yields curves with constant differences (Kay, 1977; Kalbfleisch and Prentice, 1980). Methods for the estimation of $\Lambda_s(t)$ were proposed by Cox (1972), Breslow (1972, 1974) and Kalbfleisch and Prentice (1973), and we have used Breslow's suggestion which is given by

$$\hat{\Lambda}_s(t) = \sum_{t_i \le t} \frac{\delta_i}{\sum_{j \in R(t_i)} \exp\left(\hat{\beta}' z_j\right)}$$

where $\hat{\beta}$ is the maximum partial likelihood estimator of β, z_j is the covariate vector of patient j, t_i's are the distinct survival times, $R(t_i)$ is the risk set of individuals at time t_i^- and δ_i is either 1 for death or 0 for censoring. Since $\Lambda_s(t) = -\log S_s(t)$, log cumulative hazard is also referred as $\log\big(-\log(S_s(t;\bar{z})\big)$, where the average values of k covariates of the group, $\bar{z}$, are normally used in the plotting. Plots of log cumulative hazards, e.g. by PROC LIFETEST of SAS, are the most widely used diagnostic graphs for checking the proportional hazards in survival analysis.

Another graphical method for checking the proportional hazards assumption, which is suggested by Andersen (1982), is to plot the cumulative hazard of group 1 versus that of group 2. Under proportionality, that is, if $\Lambda_1(t) = \theta\Lambda_2(t)$ holds for some constant θ, the plot of $\Lambda_1(t)$ versus $\Lambda_2(t)$ should yield approximately a straight line through the origin.

Now we will examine the two above mentioned diagnostic graphs with examples of data. Figure 3 shows plots of log cumulative hazard functions of chronic liver disease data, where each covariate of PSCORE, DIR, ALB, and also the treatment (GROUP) is dichotomized into two groups. For the chronic liver disease data, approximately constant differences over time are observed for PSCORE, DIR and ALB. But, for the covariate GROUP indicating two treatments, curves are not parallel. Also nonparallel curves for the treatment groups are observed for the examples of gastric carcinoma and thymic lymphoma data. Log cumulative hazard curves are crossing for gastric carcinoma data. For thymic lymphoma data, two treatment groups initially have the same survival experiences but later their curves are diverging.

As for another graphical method, Figure 4 provides a plot of $\Lambda_1(t)$ versus $\Lambda_2(t)$ for those three graphs, for which nonproportional hazards are suspected. Plots of two examples of gastric carcinoma and thymic lymphoma data are clearly nonlinear, and the convexity indicates the hazard ratio of the first to the second group is increasing. But for chronic liver disease data, nonlinearity is not apparent for GROUP. Sometimes graphical techniques are considered to be rather subjective and a numerical test is required to support the graphical

investigations in certain cases.

4.2. Numerical methods.

Four numerical methods of Schoenfeld (1980), Andersen (1982), Wei (1984), and Gill and Schumacher (1987) for testing the null hypothesis of proportional hazards are compared and applied to the patient data of chronic liver disease. Cox (1972) introduced a dummy time-dependent covariate as a way of checking the proportional hazards assumption. This method, however, is restricted to testing against a specific alternative and thus we will not discuss this particular testing henceforth.

Schoenfeld (1980) extended the usual chi-square goodness of fit test for the proportional hazards model. To compute a goodness of fit statistic, one first divides time axis into r intervals $[b_0, b_1), \ldots, [b_{r-1}, b_r)$ with $b_0 = 0$ and $b_r = \infty$. If there is a single covariate, the range of a covariate is subdivided into $W_1, \ldots, W_p$ sets and then the Cartesian products of an interval on the covariate axis and a time interval, $C_{sj} = W_s \times [b_{j-1}, b_j), s = 1, \ldots, p,\ j = 1, \ldots, r$ is made. Let D_j be the set of identifiers of individuals who are observed to fail in time interval $[b_{j-1}, b_j)$ and for $i \in D_j, t_i$ be the survival time of an ith individual. In each C_{sj}, d_{sj} is the observed number of deaths whose covariate values are in W_s and whose survival times are in $[b_{j-1}, b_j)$, and e_{sj} is the expected number of deaths conditional upon the risk sets at each survival time. If β were known, then

$$e_{sj} = \sum_{l \in D_j} \frac{\sum_{i \in R(t_l)} I_s(z_i) \exp\left(\beta' z_i\right)}{\sum_{i \in R(t_l)} \exp\left(\beta' z_i\right)}$$

where z_i is the covariate of ith individual and $I_s(z_i)$ is the indicator function of W_s. Usually β is estimated by the maximum partial likelihood method. An argument similar to the partial likelihood of Cox (1972, 1975) is used to derive the conditional mean of d_{sj}. If the proportional hazard assumption does not hold, then for a certain interval the effect of a covariate will be greater than others. When a single covariate taking the values 0 or 1 is tested for the

assumption of proportional hazards, the following chi-square statistic,

$$Q = \sum_{j=1}^{r} \sum_{s=1}^{2} \left(d_{sj} - e_{sj}\right)^2 / e_{sj}$$

has an asymptotic χ^2 distribution with $r-1$ degrees of freedom. This gives a slightly smaller value than the statistic of the quadratic form $(d-e)'V^{-1}(d-e)$ (Peto and Pike; 1973).

We will present the detailed calculations of goodness of fit tests for the patient data of chronic liver disease, and will briefly report the results of the other two examples in section 5.

Example. For chronic liver disease data, the goodness of fit test of Schoenfeld (1980) is applied for testing whether the covariate GROUP gives rise to the proportional hazards. For the two groups of regular EIS and episodic EIS, labelling these Group 1 and Group 2 respectively, time intervals of $[0,8)$, $[8,13)$, $[13,18)$, $[18,24)$, $[24,28)$, and $[28,\infty)$ are chosen. For $z_i = (\mathrm{PSCORE}_i, \mathrm{ALB}_i, \mathrm{DIR}_i)'$ and $\hat{\beta} = (-0.7035, 0.1176, 0.1841)$, we get the values of d_{sj} and e_{sj} in Table 1. The test statistic Q becomes 2.4635 with 5 d.f. ($p = 0.22$) which is not significant.

Table 1. Observed and expected number of deaths by Schoenfeld's method

Interval		1	2	3	4	5	6	
Group 1	d_{sj}	4	1	3	1	1	2	$n_1 = 75$
Regular EIS	e_{sj}	2.49	0.78	4.36	1.23	1.12	2.02	
Group 2	d_{sj}	2	1	8	2	2	4	$n_2 = 58$
Episodic EIS	e_{sj}	3.51	1.22	6.64	1.77	1.88	3.98	

Andersen (1982) derived a test statistic to examine the inclusion of the covariate z_{k+1} in the proportional hazards model. He adopted the approach of Kalbfleisch and Prentice (1973) assuming that the underlying hazard function $\lambda_{0s}(t)$ is constant in each time interval $[t_{j-1}, t_j)$, $j = 1, \ldots, r$, and $\lambda_{0s}(t) = \lambda_{sj}$.

The maximum likelihood estimator of λ_{sj} is

$$\hat{\lambda}_{sj} = \frac{d_{sj}}{\sum_{i=1}^{n_s} \exp\left(\beta' z_{is}\right) B_{isj}}, \quad s = 1, \ldots, p,\ j = 1, \ldots, r,$$

where n_s is the number of individuals in stratum s, B_{isj} is the time spent by i-th individual in stratum s and jth interval and z_{is} is the individual's covariate vector. The hypothesis that the covariate z_{k+1} gives rise to the proportional hazards is formulated as

$$\lambda_{s+1,j} = \lambda_{sj} \exp\left(\alpha_{s+1}\right),$$

$$\log \lambda_{s+1,j} = \log \lambda_{sj} + \alpha_{s+1} = \log \lambda_{1j} + \sum_{l=2}^{s+1} \alpha_l, \qquad s = 1, \ldots, p-1,\ j = 1, \ldots, r.$$

Using that $\log \hat{\lambda}_{sj}$, $s = 1, \ldots, p$, $j = 1, \ldots, r$, are asymptotically independent and normally distributed with mean $\log \lambda_{sj}$ and variance $\{E(d_{sj})\}^{-1}$ under the hypothesis when $n_s \to \infty$, and also using $\{d_{sj}\}^{-1}$ as an estimator of the variance, the weighted least squares estimates of the parameters, $\hat{\alpha}_{s+1}$ and $\log \hat{\lambda}_{1j}$, are

$$\hat{\alpha}_{s+1} = \frac{\sum_{j=1}^{r} \frac{\log \hat{\lambda}_{s+1,j} - \log \hat{\lambda}_{sj}}{(d_{s+1,j})^{-1} + (d_{sj})^{-1}}}{\sum_{j=1}^{r} \left((d_{s+1,j})^{-1} + (d_{sj})^{-1}\right)^{-1}} \quad s = 1, \ldots, p-1,$$

$$\log \hat{\lambda}_{1j} = \log \hat{\lambda}_j^* = \frac{\sum_{s=1}^{p} d_{sj} (\log \hat{\lambda}_{sj} - \sum_{l=1}^{s} \hat{\alpha}_l)}{\sum_{s=1}^{p} d_{sj}} \quad j = 1, \ldots, r,$$

where $\hat{\alpha}_1 \equiv 0$. From the linear models approach to the $\log \hat{\lambda}_{sj}$, the assumption of proportional hazards can then be tested by the likelihood ratio test statistic,

$$Q = \sum_{j=1}^{r} \sum_{s=1}^{p} d_{sj} \left[\log \hat{\lambda}_{sj} - \left(\log \hat{\lambda}_{1j} + \sum_{l=1}^{s} \hat{\alpha}_l\right)\right]^2 \sim \chi^2\left((r-1)(p-1)\right),$$

when $n_1, \ldots, n_p \to \infty$.

Example. Assuming the underlying hazard functions of regular EIS and episodic EIS group are constant in each interval of $[0, 8)$, $[8, 13)$, $[13, 18)$,

$[18, 24)$, $[24, 28)$, $[28, \infty)$, we obtain the values of d_{sj} and $\log \hat{\lambda}_{sj}$. $\log \hat{\lambda}_{1j}$, $\hat{\alpha}_s$ and the expected number of deaths e_{sj} are calculated under the proportional hazards assumption. The test statistic is $Q = 1.603$ with 5 d.f. $(p = 0.90)$, which is clearly not significant. Thus, the proportionality is not rejected.

Table 2. Summary statistics of Andersen's method and the number of deaths as in Table 1

Interval		1	2	3	4	5	6	
Group 1 Regular EIS	d_{sj}	4	1	3	1	1	2	$n_1 = 75$
	e_{sj}	2.26	0.75	4.14	1.13	1.13	2.26	$\hat{\alpha}_1 = 0$
	$\log \hat{\lambda}_{sj}$	-4.290	-4.733	-3.429	-4.335	-3.684	-3.773	
Group 2 Episodic EIS	d_{sj}	2	1	8	2	2	4	$n_2 = 58$
	e_{sj}	3.74	1.25	6.86	1.87	1.87	3.74	$\hat{\alpha}_2 = 0.5044$
	$\log \hat{\lambda}_{sj}$	-4.594	-4.712	-2.402	-3.682	-3.105	-3.244	
	$\log \hat{\lambda}_j^*$	-4.559	-4.975	-3.049	-4.236	-3.634	-3.757	

The advantage of the two tests of Schoenfeld (1980) and Andersen (1982) is apparent; if departures from proportional hazards are indicated with the test result, the element values of each interval may suggest just where the proportionality is failing. On the other hand, the tests of Wei (1984) and Gill and Schumacher (1987), which will be introduced shortly, do not provide these details.

Wei (1984) suggested an omnibus goodness of fit test for the proportional hazards model. Tests of Wei and Gill and Schumacher are derived from counting process models having multiplicative intensity processes. Consider two groups of sizes n_1 and n_2 and let $n = n_1 + n_2$. Let $\Lambda_s(t)$ be the cumulative hazard function of stratum s, where $s = 1, 2$. If two hazard functions are proportional, then $\Lambda_1(t) = \theta \Lambda_2(t)$ for some constant θ, which is usually called a relative risk. Wei (1984) defined $\hat{\theta}$ as the Cox's maximum partial likelihood estimator of θ, and derived a test statistic for the hypothesis of proportional hazards. For the following process

$$U_n(\theta; t) = \int_0^t dN_1(s) - \int_0^t \frac{Y_1(s)\theta}{Y_1(s)\theta + Y_2(s)} d\big(N_1(s) + N_2(s)\big)$$

$$= \sum_{k=1}^{n_1} \delta_{1k} I(t_{1k} \le t) - \sum_{s=1}^{2} \sum_{k=1}^{n_s} \frac{Y_1(t_{sk})\theta}{Y_1(t_{sk})\theta + Y_2(t_{sk})} \delta_{sk} I(t_{sk} \le t)$$

$$= N_1(t) - E_1(t),$$

where t_{sk} is a survival time of the kth individual in stratum s, $N_s(t)$ and $E_s(t)$ are respectively the observed and expected number of deaths in stratum s until time t, $Y_s(t)$ is the number of observations in $R(t)$ in stratum s, δ_{sk} is 1 if t_{sk} is a failure time or 0 if t_{sk} is a censoring time, and $I(t_{sk} \le t)$ is the identity function, $s = 1, 2$, $k = 1, \ldots, n_s$, the natural goodness of fit test based on $U_n(\hat{\theta}; t)$ is

$$T_n = \left(n\hat{\theta}\hat{\eta}(\infty)\right)^{-\frac{1}{2}} \max_{0<t<\infty} |U_n(\hat{\theta}; t)|$$

$$= \left(n\hat{\theta}\hat{\eta}(\infty)\right)^{-\frac{1}{2}} \max_{s=1,2\ k=1,\ldots,n_s} |U_n(\hat{\theta}; t_{sk})|$$

where

$$\hat{\eta}(t) = n^{-1}\left(\int_0^t \frac{Y_1(s)Y_2(s)}{\left(Y_1(s)\hat{\theta} + Y_2(s)\right)^2} d\left(N_1(s) + N_2(s)\right)\right)$$

$$= n^{-1} \sum_{s=1}^{2} \sum_{k=1}^{n_s} \frac{\delta_{sk} Y_1(t_{sk}) Y_2(t_{sk})}{\left(Y_1(t_{sk})\hat{\theta} + Y_2(t_{sk})\right)^2}.$$

Wei (1984) proved consistency of the statistic T_n against the alternative of nonproportional hazards. Under the proportional hazards, the statistic T_n converges in distribution to a process whose distribution has been extensively studied; a table published in Koziol and Byar (1975) can be used.

Example. Under the proportional hazards assumption between regular EIS and episodic EIS group, the relative risk is estimated as $\hat{\theta}$=0.5593. From Table 3, $\hat{\eta}(\infty) = 13.091/133 = 0.098$ is obtained. Since $\max_{s=1,2\ k=1,\ldots,n_s} |U_n(\hat{\theta}; t_{sk})|$ = 2.1985, the test statistic T_n yields a value of 0.8125 with a p-value 0.5.

Table 3. Summary statistics of Wei's and Gill & Schumacher's method

	Number of observations		Statistics of Wei				Statistics of Gill & Schumacher	
t	$Y_1(t)$	$Y_2(t)$	$n\hat{\eta}(t)$	$N_1(t)$	$E_1(t)$	$U_n(\hat{\theta};t)$	$K_1(t)$	$K_2(t)$
6	75	58	0.871	2	0.839	1.161	4350	32.707
7	71	57	1.731	4	2.482	1.518	4047	31.617
10	59	52	0.425	4	2.870	1.130	3068	27.639
12	58	51	0.425	5	3.259	1.741	2958	27.138
13	55	49	1.695	7	4.801	2.199	2695	25.914
14	51	45	0.425	7	5.189	1.811	2295	23.906
15	50	44	0.849	7	5.967	1.033	2200	23.404
16	49	39	1.733	8	7.617	0.383	1911	21.716
18	44	34	0.436	8	8.037	0.037	1496	19.180
19	43	32	0.438	9	8.466	0.534	1376	18.347
22	33	31	0.418	9	8.840	0.160	1023	15.984
24	32	30	0.418	9	9.213	0.213	960	15.484
25	32	29	0.422	9	9.595	0.595	928	15.213
27	27	27	0.411	10	9.954	0.046	729	13.500
28	25	27	1.607	12	11.318	0.682	675	12.981
30	20	24	0.388	12	11.636	0.364	480	10.909
39	10	11	0.399	12	11.973	0.027	110	5.238
		Total:	13.091					

Gill and Schumacher (1987) derived a test based on the estimator of the relative risk. It can be simply estimated by the generalized rank estimator

$$\hat{\theta}_{K_\theta} = \int_0^t \frac{K_0(u)}{Y_2(u)}\, dN_2(u) \Big/ \int_0^t \frac{K_0(u)}{Y_1(u)}\, dN_1(u)$$

$$= \frac{\sum_{u\le t} \frac{K_0(u)}{Y_2(u)} n_2(u)}{\sum_{u\le t} \frac{K_0(u)}{Y_1(u)} n_1(u)} = \frac{\hat{K}_{02}}{\hat{K}_{01}}$$

where the function $K_0(t)$ is a predictable random weight function and $n_s(t)$ is the number of deaths at time t in stratum s. Under the assumption of proportional hazards, the difference between $\hat{\theta}_{k_1}$ and $\hat{\theta}_{k_2}$ for two different weight functions $K_1(t)$ and $K_2(t)$, $\hat{K}_{12}/\hat{K}_{11} - \hat{K}_{22}/\hat{K}_{21}$, should be small for large sample sizes. For convenience, the symmetrized version $Q_{K_1K_2} = \hat{K}_{11}\hat{K}_{22} -$

$\hat{K}_{21}\hat{K}_{12}$, which should also be close to zero under H_0, is considered. Under H_0, $Var(Q_{K_1,K_2})$ can be estimated by

$$\widehat{Var}(Q_{K_1K_2}) = \hat{K}_{21}\hat{K}_{22}\hat{V}_{11} - \hat{K}_{11}\hat{K}_{22}\hat{V}_{21} - \hat{K}_{21}\hat{K}_{12}\hat{V}_{12} + \hat{K}_{11}\hat{K}_{12}\hat{V}_{22}$$

where

$$\begin{aligned}\hat{V}_{ii'} &= \int K_i(u)K_{i'}(u)\Big(Y_1(u)Y_2(u)\Big)^{-1}\, d\Big(N_1(u)+N_2(u)\Big) \\ &= \sum_{u\le t} K_i(u)K_{i'}(u)\Big(Y_1(u)Y_2(u)\Big)^{-1}\,\Big(n_1(u)+n_2(u)\Big).\end{aligned}$$

Then $T_{K_1K_2} = \dfrac{Q_{K_1K_2}}{\sqrt{\widehat{Var}(Q_{K_1K_2})}}$ is asymptotically a standard normal distribution as $n \to \infty$. The choice of weight functions would yield different test results.

Example. For chronic liver disease data, we choose $K_1(t)$ as the Gehan's weight function, i.e., $Y_1(t)Y_2(t)$, and $K_2(t)$ as the weight function of log rank test, i.e., $Y_1(t)Y_2(t)/\Big(Y_1(t)+Y_2(t)\Big)$. Calculations of $K_1(t)$ and $K_2(t)$ are shown in Table 3. For the Gehan's weight function we obtain $\hat{K}_{11} = 531$ and $\hat{K}_{12} = 830$, and for the log rank weight function we obtain $\hat{K}_{21} = 5.581$ and $\hat{K}_{22} = 9.977$. And the asymptotic variance of $Q_{K_1K_2}$ under H_0 is estimated as $\widehat{Var}(Q_{K_1K_2}) = 367203.04$. Therefore, based on a Gehan versus log rank comparison, a standardized test statistic is $T_{K_1K_2} = 1.09(p = 0.28)$, and does not reject the proportionality.

5. EXAMPLES

5.1. Medical data.

There is a simple indicator of nonproportionality in applications; the log rank and Gehan's generalized Wilcoxon tests might give different answers particularly in nonproportional situations (Gill and Schumacher, 1987). P-values of the GROUP variable for chronic liver disease data are 0.109 for the log rank and 0.239 for Wilcoxon test. And p-values of the gastric carcinoma data are 0.312 for the log rank and 0.030 for Wilcoxon test. Likewise different p-values for thymic lymphoma data are 0.070 for log rank and 0.380 for Wilcoxon test.

We now apply the four goodness of fit tests to these data sets, and the results are summarized in Table 4. For chronic liver disease data, all four tests give different p-values but all of them are insignificant. The test results of Schoenfeld (1980) and Andersen (1982) differ depending on the partitioning of time intervals.

For the gastric carcinoma data with crossing hazards, all four tests reject the null hypothesis. However, for the thymic lymphoma data with diverging hazards, only the test of Gill and Schumacher (1987), based on a log rank versus Gehan comparison, definitely rejects the null hypothesis and the other tests provide a borderline significance. Again Schoenfeld's test provides diverse results with different time intervals. As the pattern of diverging hazards is clearly demonstrated from 240 days onwards, it is promising that at least any of the four tests rejects the null hypothesis. However, it should be reminded that Gill and Schumacher's test is not a guarantee for all situations. Gill and Schumacher (1987) presented a data set which failed to reject the null hypothesis despite of its apparent nonproportionality.

Table 4. Performances of the four goodness of fit tests

	Schoenfeld	Andersen	Wei	Gill & Schumacher
Chronic liver disease (GROUP)	$\chi^2 = 2.300$ $d.f. = 2$ $p = 0.317$	$Q = 0.577$ $d.f. = 2$ $p = 0.749$	$T = 0.813$ $p = 0.45$–0.5	$T = 1.098$ $p = 0.27$
	$\chi^2 = 2.464$ $d.f. = 5$ $p = 0.220$	$Q = 1.603$ $d.f. = 5$ $p = 0.901$		
Gastric carcinoma	$\chi^2 = 18.268$ $d.f. = 5$ $p = 0.003$	$Q = 12.645$ $d.f. = 5$ $p = 0.027$	$T = 1.953$ $p \leq 0.0001$	$T = 3.872$ $p = 0.00005$
Thymic lymphoma	$\chi^2 = 5.761$ $d.f. = 3$ $p = 0.124$	$Q = 6.463$ $d.f. = 3$ $p = 0.091$	$T = 1.29$ $p = 0.06*$	$T = -2.247$ $p = 0.025$
	$\chi^2 = 9.561$ $d.f. = 4$ $p = 0.049$	$Q = 8.754$ $d.f. = 4$ $p = 0.068$		

* The result was provided by Wei (1984).

5.2. Simulations.

The performance of the tests is studied using Monte Carlo simulations. The random numbers are generated by the IMSL subroutines and all simulations include 1000 samples. The survival distributions include Weibull $\big(W(\lambda,\alpha)\big)$ with density function $\alpha\lambda(\lambda t)^{\alpha-1}\exp\big(-(\lambda t)^{\alpha}\big)$, where λ, $\alpha > 0$, and exponential $\big(W(\lambda,1)\big)$ with density function $\lambda\exp(-\lambda x)$, where $\lambda > 0$. A uniform random censorship is modeled and censoring of 20% is generated approximately. Sixty observations are allocated to each group and time intervals of six are chosen. The power is determined by the percentage that the computed statistics exceed the cut off value corresponding to $\alpha = 0.05$ in 1000 trials. Two heterogeneous distributions are generated for the power calculations, and the scale parameter of a Weibull distribution is fixed to be 1.0.

The results of simulations are summarized in Table 5. The empirical sizes of tests are designated with a star, which correspond to testing the null hypothesis of proportional hazards. The size of Gill and Schumacher's statistic is near to or somewhat larger than the nominal level $\alpha = 0.05$, but the sizes of other statistics are typically smaller than 0.05. Monte Carlo estimates of power of 0.05 level tests differ for the four statistics. Note that the simulation data sets have a monotone hazard ratio except the last two cases. Across a broad range of alternatives of monotone departures from a constant hazard ratio, superior performance of Gill and Schumacher's test is demonstrated. And Wei's statistic is the next most powerful and Andersen's statistic is slightly less powerful than that of Schoenfeld. Nonmonotone changes of the hazard ratio through time are artificially produced and powers of these alternatives clearly demonstrate that Gill and Schumacher's test is not a best choice, and tests of Schoenfeld and Andersen are rather safe in situations of nonmonotone hazard ratio. Fluctuating behavior of Wei's statistic is noticed.

Table 5. Empirical powers of the four goodness of fit tests

Distributions		Tests of			
Group 1	Group 2	Schoenfeld	Andersen	Wei	Gill& Schumacher
$*W(0.5,1)$	$W(0.5,1)$	0.035	0.029	0.030	0.042
$*W(0.5,1)$	$W(2.0,1)$	0.036	0.032	0.039	0.050
$*W(1,0.5)$	$W(1,0.5)$	0.032	0.062	0.045	0.073
$W(1,0.5)$	$W(1,0.7)$	0.213	0.230	0.324	0.468
$*W(1,2.0)$	$W(1,2.0)$	0.040	0.023	0.049	0.067
$W(1,2.0)$	$W(1,2.2)$	0.060	0.045	0.059	0.086
$W(1,2.0)$	$W(1,2.4)$	0.065	0.041	0.119	0.206
$W(1,2.0)$	$W(1,2.6)$	0.120	0.069	0.186	0.334
$W(1,2.0)$	$W(1,2.8)$	0.194	0.128	0.338	0.495
$W(1,2.0)$	$W(1,3.0)$	0.312	0.225	0.479	0.640
$W(1,2.0)$	$W(1,4.0)$	0.784	0.695	0.890	0.962
$W(1,2.0)$	**	0.313	0.271	0.142	0.030
$W(1,2.0)$	**	0.834	0.749	0.913	0.029

* Monte Carlo estimates of size of tests

** Heterogeneous distributions, based on the Weibull with a scale parameter 1.

There are several weaknesses in this simulation. Tests of Schoenfeld and Andersen depend on the partitioning of time intervals and each interval has to contain a reasonable number of deaths. If the generated observation times do not produce a failure within any of six intervals, we abandoned that sample, so the powers of tests of Schoenfeld and Andersen are calculated based on less than 1000 iterations. The effect of number of intervals on the performance of Schoenfeld's and Andersen's test, and the effect of the extent of censoring on the four tests need to be studied in the future.

6. DISCUSSION

There is an increasing awareness of the adverse effects of model misspecification on the statistical inference, and thus it is advised for the users of Cox models to perform goodness of fit analysis. The present paper has shown that the proportionality of hazards assumption can be investigated by graphical and numerical methods. In our study, plotting the estimate of integrated

baseline hazard function of one group against another is more useful for checking proportionality than plotting the log integrated hazard function against time. The graphical examinations can be reinforced by the numerical goodness of fit tests for the proportionality of hazards. Advantages of the four tests differed. Although the tests of Schoenfeld (1980) and Andersen (1982) provided details of where the proportionality might fail, different test results might be obtained depending on different partitioning of time intervals. Assuming constant underlying hazards within each time interval by Andersen's test deviates somewhat from the idea of arbitrary hazards which was originally assumed by Cox model. However, if the underlying hazard is a slowly varying function of time, then the hazards within each interval can be assumed to be constant. Although Gill and Schumacher's test performs best under monotone hazard ratio alternative, it performs worst under nonmonotone hazard ratio alternative. Hence, if a more precise form of the alternative can be considered before proceeding to a test, it would help us choose an appropriate test for a problem.

When this study was completed, we were introduced to a paper of Lin and Wei (1991) by a personal correspondence with one of the authors. Their paper has an added feature of extending the information matrix test, originally proposed by White (1982), as a goodness of fit test for the proportional hazards assumption. A part of simulation results by Lin and Wei is conformable with ours.

Figure 1. Estimated Kaplan-Meier survival functions for two treatment groups of chronic liver disease data: solid line for regular EIS group and dashed line for episodic EIS group.

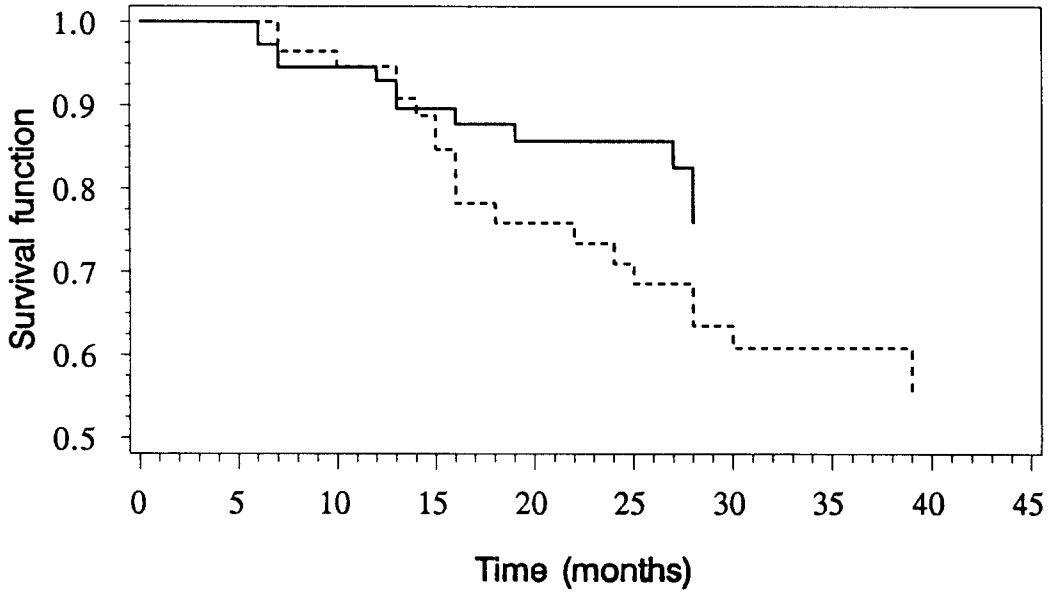


Figure 2. Estimated hazard functions for two treatment groups of chronic liver disease data : lines are the same as in Fig. 1.

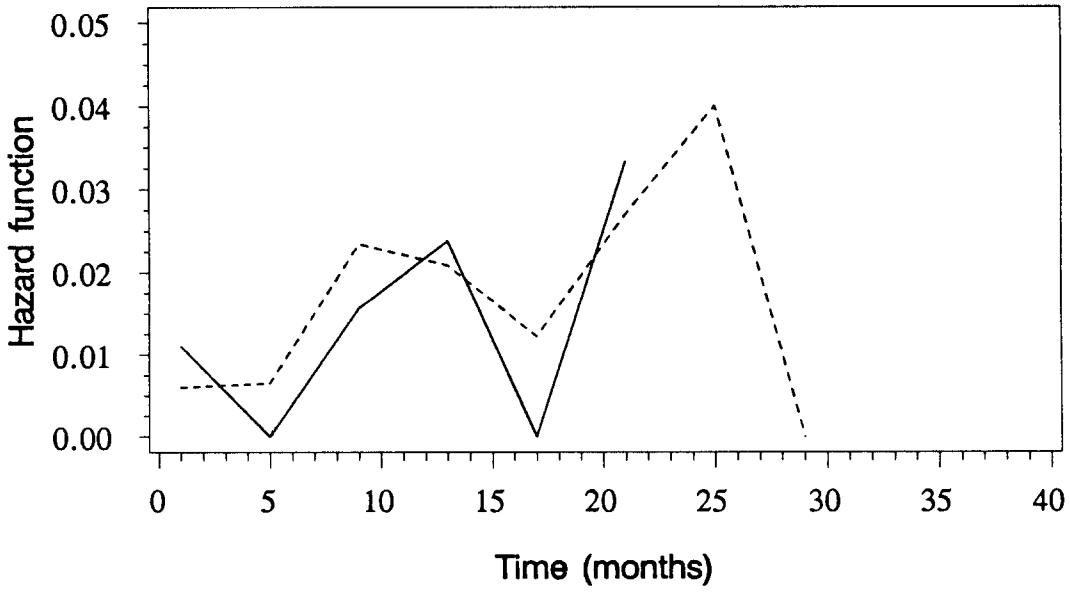


Figure 3. Estimated log cumulative hazard functions for chronic liver disease, gastric carcinoma and thymic lymphoma data.

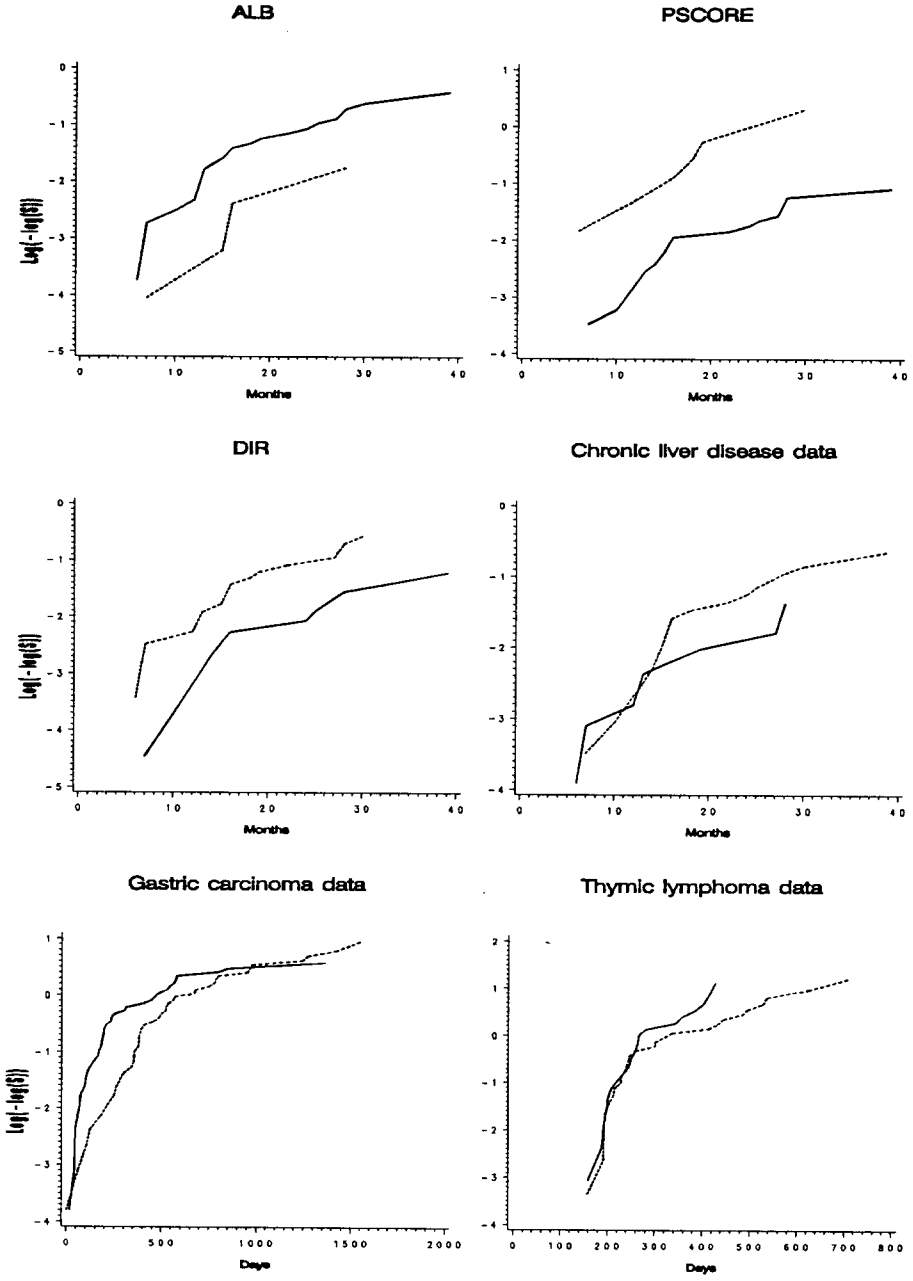


Figure 1. Estimated Kaplan-Meier survival functions for two treatment groups of chronic liver disease data: solid line for regular EIS group and dashed line for episodic EIS group.

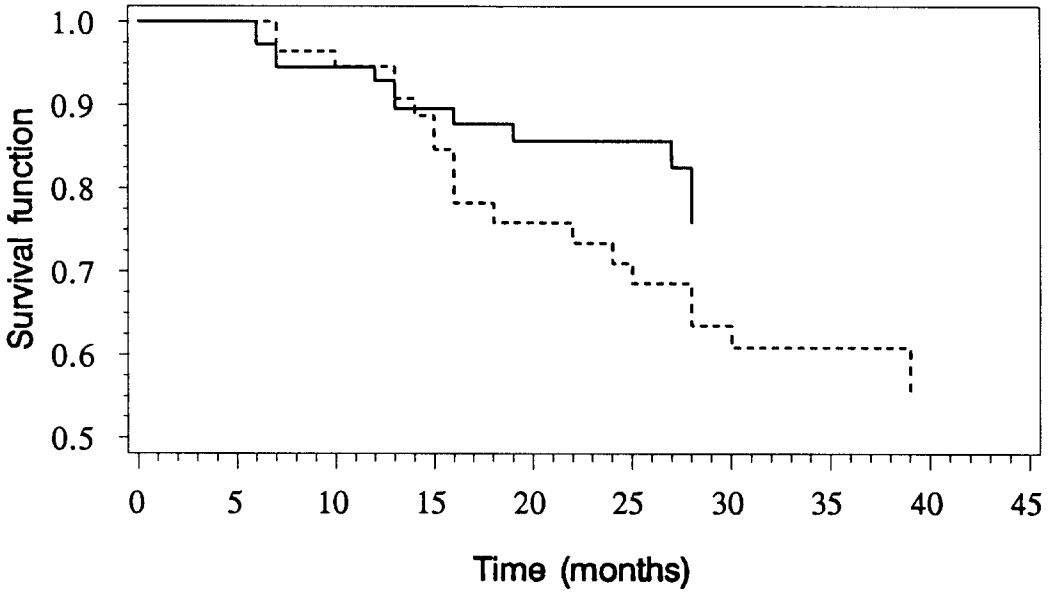


Figure 2. Estimated hazard functions for two treatment groups of chronic liver disease data : lines are the same as in Fig. 1.

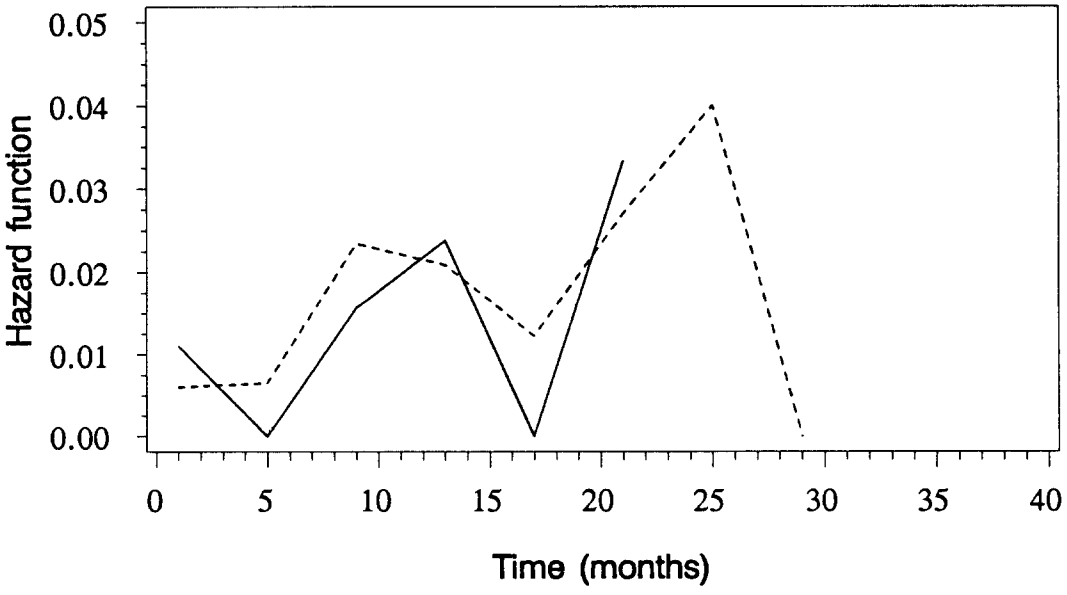


Figure 3. Estimated log cumulative hazard functions for chronic liver disease, gastric carcinoma and thymic lymphoma data.

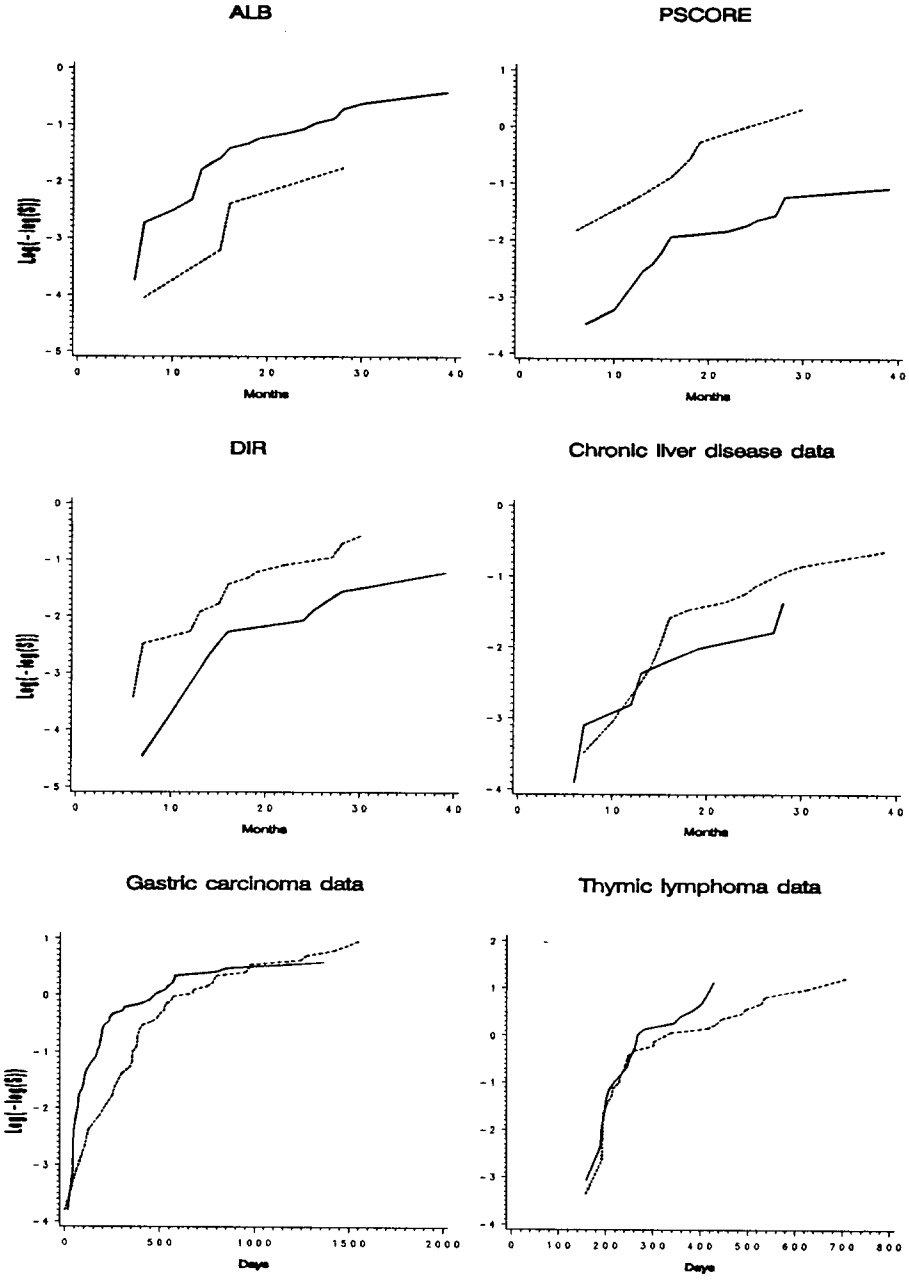
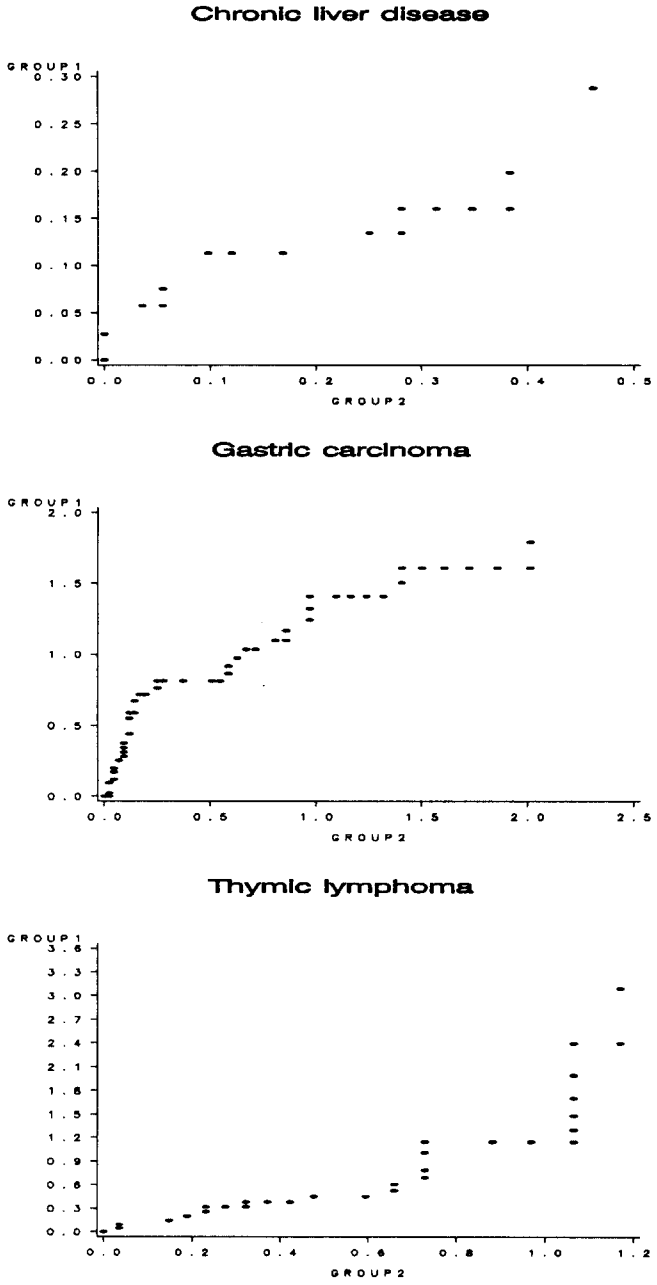


Figure 4. Estimated cumulative hazard function of group 1 versus that of group 2 for chronic liver disease, gastric carcinoma and thymic lymphoma data.



REFERENCES

(1) Andersen, P.K. (1982). Testing goodness of fit of Cox's regression and life model. *Biometrics*, Vol.**38**, **67–77**.

(2) Andersen, P.K. (1983). Comparing survival distributions via hazard ratio estimates. *Scandinavian Journal of Statistics*, Vol.**10**, **77–85**.

(3) Anderson, J. and Senthilselvan, A. (1982). A two step regression model for hazard functions. *Applied Statistics*, Vol.**31**, **44–51**.

(4) Begun, J.M. and Reid, N. (1983). Estimating the relative risk with censored data. *Journal of the American Statistical Association*, Vol.**78**, **337–341**.

(5) Breslow, N. (1972). Discussion on professor Cox's paper. *Journal of Royal Statistical Society*, Series **B**, Vol.**34**, **216–217**.

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

(7) Carter, W.H., Wampler, G.L., and Stablein, D.M. (1983). *Regression Analysis of Survival Data in Cancer Chemotherapy*, Marcel Dekker, New York.

(8) Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of Royal Statistical Society*, Vol.**34**, **187–220**.

(9) Cox, D.R. (1979). A note on the graphical analysis of survival data. *Biometrika*, Vol.**66**, **188–190**.

(10) Gill, R.D. and Schumacher, M. (1987). A simple test of the proportional assumption. *Biometrika*, Vol.**74**, **289–300**.

(11) Gore, S.M., Pocock, S.J., and Kerr, G.R. (1984). Regression models and nonproportional hazards in the analysis of breast cancer survival. *Applied Statistics*, Vol.**33**, **176–195**.

(12) Harris, E.K. and Albert, A. (1991). *Survivorship Analysis for Clinical Studies*, Marcel Dekker, New York.

(13) Hoel, D.G. (1972). A representation of mortality data by competing risks. *Biometrics*, Vol.**28**, **475–488**.

(14) Kalbfleisch, J. and prentice R. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, Vol.**60**, **267–278**.

(15) Kalbfleisch, J. and prentice R. (1980). *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, New York.

(16) Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, Vol.**53**, **457–481**.

(17) Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics*, Vol.**26**, **227–237**.

(18) Kay, R. (1984). Goodness of fit methods for the proportional hazards regression model: A review. *Revue de Epidemiologie et de Sante Publique*, Vol.**32**, **185–198**.

(19) Koziol, J. and Byar, D. (1975). Percentage points of the asymptotic distributions of one and two sample K-S statistics for truncated or censored data. *Technometrics*, Vol.**17**, **507–510**.

(20) Lin, D.Y. and Wei, L.J. (1991). Goodness of fit tests for the general Cox regression model. *Statistica Sinica*, Vol.**1**, **1–17**.

(21) Moreau, T., O'Quigley, J., and Mesbah M. (1985). A global goodness of fit statistic for the proportional hazards model. *Applied Statistics*, Vol.**34**, **212–218**.

(22) Pepe, M.S. and Flemming, T.R. (1989). Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics*, Vol.**45**, **497–507**.

(23) Peto, R. and Pike, M.C. (1973). Conservatism of the approximation $(o-e)^2/e$ in the log rank test for survival data or tumor incidence data. *Biometrics*, Vol.**29**, **579–584**.

(24) Peto, R., Pike, M.C., and others (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *British Journal of Cancer*, Vol.**35**, **1–39**.

(25) Pugh, R.N., Murray-Lyon, I.M., Dawson, J.L., and others (1973). Transection of the oesophagus for bleeding oesophageal varices. *British Journal of Surgery*, Vol.**60**, **646–649**.

(26) Schoenfeld, D. (1980). Chi-squared goodness of fit tests for the proportional hazards regression model. *Biometrika*, Vol.**67**, **147–153**.

(27) Terblanche, J. (1990). Has sclerotheraphy altered the management of patients with variceal bleeding? *American Journal of Surgery*, Vol.**160**, **37–42**.

(28) Terblanche, J., Northover, J., and others. (1979). A prospective controlled trial of sclerotheraphy in the long term management of patients after esophageal variceal bleeding. *Surgery, Gynecology & Obstetrics*, Vol.**148**, **323–333**.

(29) Wei, L.J. (1984). Testing goodness of fit for proportional hazards model with censored observations. *Journal of the American Statistical Association*, Vol.**79**, **649–652**.

(30) White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, Vol.**50**, **1–25**.