Annu. Rev. Public Health. 1997. 18:105–34

SURVIVAL ANALYSIS IN PUBLIC HEALTH RESEARCH

Elisa T. Lee and Oscar T. Go
College of Public Health, University of Oklahoma Health Sciences Center, P.O. Box 26901, Oklahoma City, Oklahoma 73190; e-mail: elisa-lee@uokhsc.edu

KEY WORDS: survival data, censored, Kaplan-Meier, Cox, proportional hazards

ABSTRACT

This paper reviews the common statistical techniques employed to analyze survival data in public health research. Due to the presence of censoring, the data are not amenable to the usual method of analysis. The improvement in statistical computing and wide accessibility of personal computers led to the rapid development and popularity of nonparametric over parametric procedures. The former required less stringent conditions. But, if the assumptions for parametric methods hold, the resulting estimates have smaller standard errors and are easier to interpret. Nonparametric techniques include the Kaplan-Meier method for estimating the survival function and the Cox proportional hazards model to identify risk factors and to obtain adjusted risk ratios. In cases where the assumption of proportional hazards is not tenable, the data can be stratified and a model fitted with different baseline functions in each stratum. Parametric modeling such as the accelerated failure time model also may be used. Hazard functions for the exponential, Weibull, gamma, Gompertz, lognormal, and log-logistic distributions are described. Examples from published literature are given to illustrate the various methods. The paper is intended for public health professionals who are interested in survival data analysis.

INTRODUCTION

The origins of survival analysis may be traced back to early work on mortality in the seventeenth century when Graunt published the first Weekly Bill of Mortality in London and Healey published the first lifetable. Since then, the lifetable method has been used frequently by actuaries, statisticians, and biomedical researchers in governmental and private agencies. During World War II, reliability of military equipment became a critical issue. This led to the

0163-7525/97/0510-0105$08.00

study of the durability or the "lifetime" of industrial devices. After the war, the methods used to analyze the reliability of industrial devices were further developed and applied to the study of survival time of cancer patients. The term "lifetime analysis" used by industrial reliability engineers was changed to "survival analysis" by cancer researchers. In the past four decades, survival analysis has become one of the most frequently used methods for analyzing data in disciplines ranging from medicine, epidemiology, and environmental health, to criminology, marketing, and astronomy.

The primary variable in survival analysis is survival time, no longer limited to mean the time to death. The term "survival time" is used loosely for the time period from a starting time point to the occurrence of a certain event. Examples of survival time are: the time to the development of diabetic retinopathy from the time of diagnosis of diabetes, the time to parole of a prisoner, the duration of first marriage, workman's compensation or other insurance claims, and lifetime of electronic devices and computer components.

Special Features of Survival Time

Survival time is a nonnegative random variable measuring the time interval from an origin to the occurrence of a given event. Because of its two special features, survival times are not amenable to standard statistical methods. First, the exact survival time may be longer than the duration of the study time (or observation time) and is therefore unknown. For example, in a ten-year longitudinal study of diabetic retinopathy, many patients may not develop retinopathy in the ten-year period and, therefore, the exact retinopathy-free time is unknown. If a patient without retinopathy enters the study at the beginning of year 2 and is still retinopathy-free at the end of the study, his retinopathy-free time is at least nine years (usually denoted by 9+ years). In an insurance claim, the total amount of money an insurance company pays in the case of a severe automobile accident at a certain time may not be exactly known. It may be known that the total amount is at least \$5000 (or \$5000+). In many clinical and epidemiological studies, participants may leave the study before it ends and therefore become lost to follow-up. They may die of a cause unrelated to the disease under study or simply refuse to continue their participation. For example, in a three-year study of mortality due to lung cancer, a participant may refuse to continue after two years. His survival time is 2+ years. These incomplete observations are called "censored" observation, and in particular, "right censored" observations in contrast to "left censored" observations, in which the exact survival time is unknown but is less than the observation time. Most survival analysis methods focus on right censoring since it occurs far more frequently than left censoring. The incompleteness of data makes the conventional statistical methods inappropriate.

The second feature is that survival times often follow a skewed distribution, far from normal. Although transformations can be applied to make the distribution more symmetrical, using a different model may be more satisfactory. Some of the distributions known to be appropriate for survival times include the exponential, Weibull, lognormal, gamma, Gompertz, and log-logistic.

Three Survival Functions

The distribution of survival times is characterized by three functions: (*a*) the survivorship function, (*b*) the probability density function, and (*c*) the hazard function.

Let T denote the random variable, survival time. The survivorship function, denoted by $S(t)$, is defined as $S(t) = P(T > t) = 1 - F(t)$, where $F(t)$ is the distribution function of T. $S(t) = 1$ at $t = 0$ and $S(t) = 0$ at $t = \infty$. The graph of $S(t)$ is known as the survival curve, which begins at $S(0) = 1$ and decreases to 0 as t increases to infinity.

The probability density function is defined as $f(t) = \lim_{\Delta t \to 0} P(t < T < t + \Delta t)/\Delta t$, which is also known as the *unconditional failure rate*. The hazard function, on the other hand, gives the *conditional failure rate*. It is defined as $h(t) = \lim_{\Delta t \to 0} P(t < T < t + \Delta t / T > t)/\Delta t$. The hazard function is also known as the *instantaneous failure rate, age-specific failure rate, or conditional mortality rate*. It is a measure of the proneness to failure as a function of the age of the individual in the sense that the quantity $\Delta t\, h(t)$ is the expected proportion of age t individuals who will fail in the short interval t to $t + \Delta t$. It is not really a probability since its value can be greater than one. The hazard function plays an important role in survival analysis.

The three functions are mathematically equivalent. If one of them is known, the other two can be derived. For practical purposes, the survivorship function is most useful because it gives the median survival time and other summary statistics. It is often referred to as the survival function.

Applications in Public Health

Major applications of survival analysis in public health include the following: (*a*) estimation of survival distributions, (*b*) testing hypotheses of equal survival distributions, and (*c*) identification of risk or prognostic factors. The approach taken may be nonparametric, parametric, or semiparametric.

Regardless of the approach selected, the estimation of survival distribution provides estimates of descriptive statistics such as the median survival time and the probability of surviving longer than a given period of time. In order to compare the survival distribution of two or more groups, two-sample and K-sample tests have been developed. A typical example is to compare the disease-free time of males to that of females or to compare the survival distributions of

patients in three different ethnic groups. In many studies, such comparisons are performed as a first step to the identification of important risk or prognostic factors. In addition to studying each potential risk/prognostic factor individually, simultaneous analysis of multiple factors is necessary to sort out the interrelationship of risk factors. Regression analysis, using parametric or semiparametric approaches, is usually performed.

In the following two sections, we review the most commonly used survival analysis methods for the three major applications using nonparametric/semiparametric and parametric approaches. Examples taken from the literature are used for illustrative purposes. The final section discusses briefly the available computer software for survival analysis. The paper is intended to be a review or tutorial for public health professionals who are interested in the subject.

NONPARAMETRIC AND SEMIPARAMETRIC APPROACHES

The Kaplan-Meier Product Limit Method

The Kaplan-Meier product limit (PL) method is a special case of the lifetable technique, in that a series of time intervals is formed in such a way that only one death occurs in each interval, and the death occurs at the beginning of the interval. It estimates the probability of surviving longer than a given time t, i.e. $\mathrm{S}(t)$. The estimate is the product of a series of estimated conditional probabilities. For example, the probability of surviving longer than k years is estimated as,

$$\hat{\mathrm{P}}(\mathrm{T} > k) = \hat{\mathrm{S}}(k) = p_1 \cdot p_2 \cdot p_3 \cdots p_{\mathrm{k}},$$

where p_1 denotes the proportion of patients surviving at least one year, p_2 denotes the proportion of patients surviving the second year after they have survived the first year, p_3 denotes the proportion of patients surviving the third year after they have survived the second year, and p_{k} denotes the proportion of patients surviving the kth year after they have survived $\mathrm{k}-1$ years. Let t_i represent the ith survival time, censored or uncensored and δ_{i} be an indicator variable with a value of 0 for censored observations and 1 for uncensored observations. Now rank the values of t_i in ascending order of magnitude. If $\mathrm{r_i}$ is the rank of t_i, the Kaplan-Meier estimate may be written as

$$\hat{\mathrm{S}}(t) = \prod_{\mathrm{t_i}<\mathrm{t}} \left(\frac{\mathrm{n}-\mathrm{r_i}}{\mathrm{n}-\mathrm{r_i}+1} \right)^{\delta_{\mathrm{i}}}, \mathrm{t} \le \mathrm{t}_{(\mathrm{n})}$$

where n is the total number of observations and $t_{(n)}$ the largest observed survival time. When the estimated median survival time is the 50th percentile, the value

of t at $\hat{S}(t) = 0.5$. An important assumption of the Kaplan-Meier method is that the probability of a censored observation is independent of the actual survival time. In other words, the reason an observation is censored is unrelated to the cause of death under study, and the individuals whose survival times happen to be censored represent a sample from the same distribution as the others.

The Kaplan-Meier method is useful in estimating the survival distribution, $S(t)$. However, the PL estimates are limited to the time interval in which the observations fall. If the largest observation is uncensored, the PL estimate at that time is always zero. If the largest observation is censored, the PL estimate can never equal zero and is undefined beyond the largest observation, unless an additional assumption is imposed. In addition, if less than 50% of the observations are uncensored and the largest observation is censored, the median survival time cannot be estimated. Thus, the method is not perfect and there are reasons to search for a parametric model.

The Kaplan-Meier method is commonly used by medical researchers and epidemiologists. Examples can easily be found in the literature. Recently, it was applied to health economics by Fenn et al (17). The authors advocate the use of survival analysis, particularly the Kaplan-Meier method in the economic evaluation of cost-effectiveness of treatment where censored cost data are present. Censored data arise when the course of treatment extends beyond the end of the clinical trial period and when patients withdraw from the trial for reasons unconnected with the treatment under study.

The Kaplan-Meier method was also applied to health-related vocational management by Saeki et al (49) in a retrospective cohort study to evaluate the longitudinal trend of proportion of patients who return to work after their first stroke. The event of interest is the "return to work after first stroke," which was defined as one month or more of work in active employment after stroke. Thus, patients who were followed up but did not return to work, or patients who were lost to follow-up, were censored cases. A total of 183 stroke patients discharged alive from the University of Occupational and Environmental Health Hospital in Kitakyushu, Japan, were included in the study. The patients were 18–64 years of age and had competitive employment status at the time of stroke. In addition to time to return to work, the authors also investigated possible predictors such as normal muscle strength, absence of apraxia, and occupation.

Figure 1 shows the proportion of return to work after stroke by the Kaplan-Meier method. Note that the authors chose to plot $F(t) = 1 - S(t)$ versus t. Two steep slopes were found in the curve. The first slope in the first six months after admission was considered due to early discharge from the hospital. The second slope occurred from 12 to 18 months after admission. The authors stated that it seemed to coincide with the expiration of patient sickness benefits from the public fund source (generally limited to 18 months in Japan). The authors

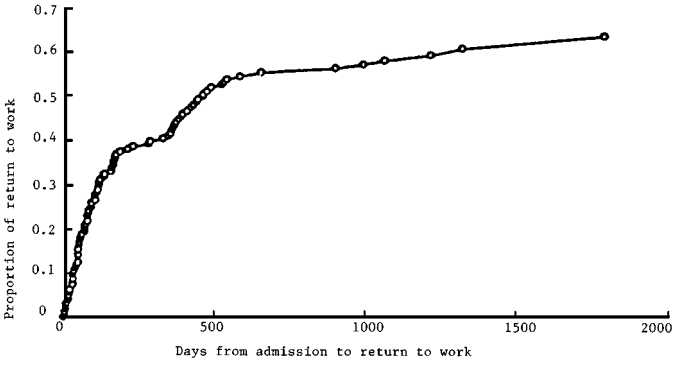

Figure 1 Graph shows proportion of return to work after stroke in Kaplan-Meier method (n = 183). Source: Saeki et al, 1995. Reproduced with permission from *Stroke*.

concluded that stroke patients either return to work early after stroke or prefer to receive longer sickness benefits and delay their time to return to work. The authors also estimated and plotted the proportion of return to work after stroke by muscle strength, apraxia, and occupation (Figures 2–4) as a first attempt to identify predictors (more discussion is given in the following sections).

Testing the Equality of Survival Distributions

Comparisons of survival distributions are common in public health research. For example, a diabetes epidemiologist may want to compare the retinopathy-free times of patients who have hypertension to patients with normal blood pressure to see if hypertension is related to the development of retinopathy.

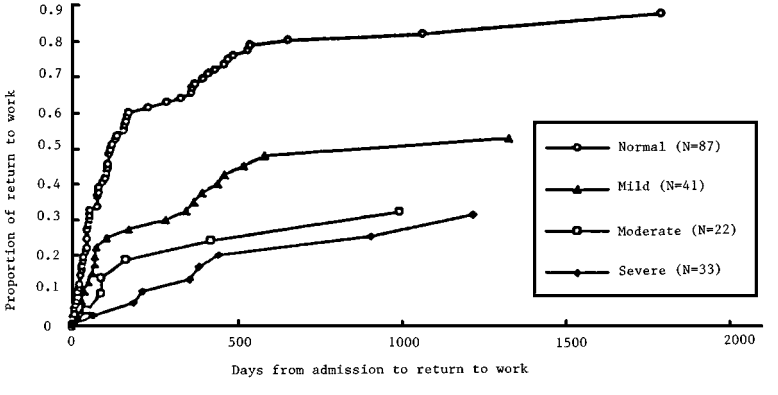

Figure 2 Graph shows proportion of return to work after stroke by muscle strength in Kaplan-Meier method (n = 183). Curves differ by log-rank test ($P = .0001$). Source: Saeki et al, 1995. Reproduced with permission from *Stroke*.

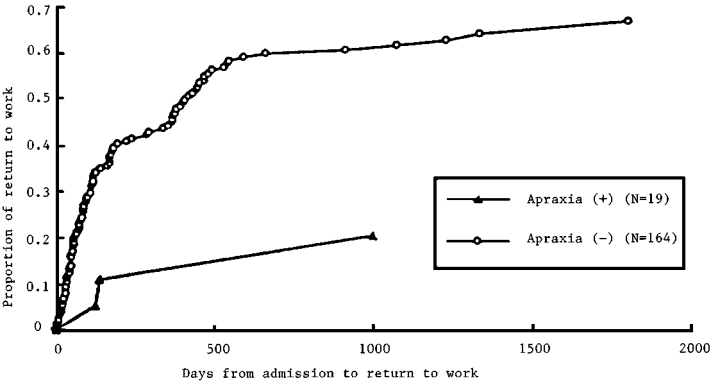

Figure 3 Graph shows proportion of return to work after stroke by apraxia in Kaplan-Meier method (n = 183). Curves differ by log-rank test ($P = .0010$). Source: Saeki et al, 1995. Reproduced with permission from *Stroke*.

A health educator may wish to study three smoking cessation strategies and compare the time to restarting among participants who have undergone the three interventions in order to identify the best strategy.

Two commonly used nonparametric tests for comparison of two survival distributions are the Gehan's Generalized Wilcoxon test (19, 20) and the logrank test (41). Both tests are based on the ranks of the observations. The Mantel's (34) version of the Wilcoxon test pools the two samples of size n_1 and n_2. Assign

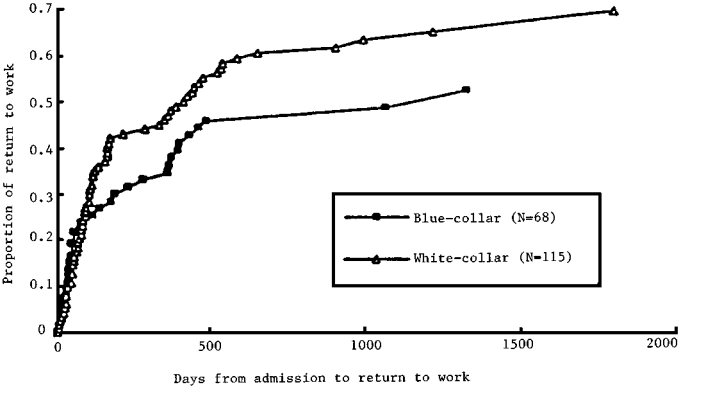

Figure 4 Graph shows proportion of return to work after stroke by occupation in Kaplan-Meier method (n = 183). Curves are not significantly different by log-rank test ($P = 0981$), but significantly different by Wilcoxon's test ($P = .0492$). Source: Saeki et al, 1995. Reproduced with permission from *Stroke*.

a score to each observation based on its relative ranking in the combined set. Let U_i, $i = 1, 2, \ldots, n_1 + n_2$ be the number of remaining $n_1 + n_2 - 1$ observations that the ith observation is definitely greater than minus the number that it is definitely less than. The test statistic, W, of Gehan's Wilcoxon test and its variance are

$$W = \sum_{i=1}^{n_1} U_i^2, \quad \mathrm{Var}(W) = \frac{n_1 n_2 \sum_{i=1}^{n_1+n_2} U_i^2}{(n_1 + n_2)(n_1 + n_2 - 1)}.$$

and $W/\sqrt{\mathrm{Var}(W)}$ is asymptotically normal with mean 0 and variance 1. An appropriate critical region of testing whether there is any significant difference between the two survival functions can be constructed from the standard normal distribution.

Let O_1 and O_2 be the observed numbers and E_1 and E_2 the expected numbers of deaths in the two groups. The logrank test statistic is equivalent to

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2},$$

which has approximately the chi-square distribution with one degree of freedom. A large X^2 value would lead to the rejection of the null hypothesis in favor of the alternative that the two treatments are not equally effective.

The logrank test is appropriate for survival distributions whose hazard functions are proportional over time, i.e. the two survival curves do not cross. Otherwise, the Gehan's Wilcoxon test is recommended. The Wilcoxon statistic puts more weight on early deaths compared to the logrank.

For comparison of three or more survival distributions with censoring, an extension of the Kruskal-Wallis test is applicable. The k-sample test for censored data (41) is also based on rank statistics. Let n_j be the number of observations in the jth sample and $N = \sum_{j=1}^{k} n_j$. Pool the k samples and obtain a set of scores w_i, $i = 1, \ldots, N$ according to a scoring method such as that in Gehan's Wilcoxon test. The test statistics is

$$X^2 = \frac{\sum_{j=1}^{k} \frac{S_j^2}{n_j}}{s^2},$$

where S_j is the sum of the scores in the jth sample and $s^2 = \sum_{i=1}^{N} w_i^2/(N-1)$. Under the null hypothesis X^2 has approximately chi-square distribution with $k - 1$ degrees of freedom.

Applications of these two- and k-sample tests easily can be found in public health journals. For example, in the study of return to work after stroke of Saeki et al (49), patients were divided into subgroups according to the level of impaired muscle strength, presence of apraxia, and occupation (Figures 2–4).

The Wilcoxon test and logrank test were applied to compare the distributions of "time from hospital admission to return to work after stroke" of different subgroups. The proportion of return to work after stroke was significantly different ($p = 0.0001$) among the four muscle strength groups with the higher proportion in the group of patients with less impaired muscle strength. Patients with normal muscle strength returned to work much sooner than those with muscle strength impairment. Patients without apraxia also returned to work much earlier than those with apraxia ($p = 0.001$). The proportion of patients who returned to work after stroke was higher in blue-collar workers than white-collar workers in the first three months. After the first three months, the white-collar workers had a higher rate of return to work. The difference between the two groups was found to be significant by the Wilcoxon test ($p = 0.049$), but not by the logrank test ($p = 0.098$).

Cox's Regression Model

In public health research, it is often of interest to know whether a certain personal characteristic is related to the occurrence of a certain health-related event. For example, are cigarette smoking, elevated cholesterol value, and family history of heart disease related to the development of cardiovascular disease? In this case, cigarette smoking, cholesterol value, and family history of heart disease are referred to as risk factors (or in a clinical setting, prognostic factors), or covariates. In most public health studies, data on many risk factors are collected and therefore the identification of the most significant risk factors becomes an important task.

In addition to examining individually each variable's relationship to the length of disease-free time, survival, or remission, multivariate regression analysis is necessary to control for confounding factors. Cox's (9) regression model has been the most widely used method in survival data analysis regardless of whether the survival time is discrete or continuous and whether there is censoring.

Given a set of p covariates $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$, the hazard function for a given individual i is modeled by

$$h_i(t, \mathbf{x}_i) = h_0(t) \exp(\mathbf{b}^T \mathbf{x}_i). \qquad 1.$$

That is, the hazard is the product of an arbitrary nonnegative baseline hazard $h_0(t)$ and an exponential linear function of the p covariates. The linear function $\mathbf{b}^T \mathbf{x}_i = b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi}$ also is known as the prognostic index for the *i*th individual which is exponentiated to insure a nonnegative hazard. One feature of the model is the absence of the intercept term. Since the distribution of the survival times is not specified, the Cox model is an example of a semi-parametric model. Taking the natural logarithm of Equation 1 and suppressing

the subscripts i gives

$$\log_e h(t, \mathbf{x}) = \log_e h_0(t) + \mathbf{b}^T\mathbf{x}. \quad 2.$$

For this setup, the hazard ratio of any two individuals with covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by $\exp[\mathbf{b}^T(\mathbf{x}_1 - \mathbf{x}_2)]$, which does not depend on the baseline hazard and is constant over time. In other words, the two hazard functions do not cross. This property is referred to as proportional hazards and is also shared by some parametric models to be discussed in the latter sections.

Though the Cox model is introduced initially in the framework of proportional hazards, the model easily can be extended to cases of nonproportional hazard functions. Some examples are models that include covariates that are time dependent (i.e. their values change over the follow-up period) and those that have multiple events for some individuals.

In addition to being a regression model, Equation 1 can be applied to two-sample problems with and without covariates. Consider the case of comparing survival times of two groups of patients. Let $p = 1$ and x_1 be the treatment indicator variable that takes the value 0 if a patient is in group 1 and 1 if in group 2. The proportional hazards assumption implies that the hazard function for group 2 is proportional to that of the hazard for group 1. This is written as $h_2(t) = e^b h_1(t)$. If $e^b > 1$, the survival of group 1 is superior to that of group 2. This reduces to Equation 1 with $h_1(t)$ equals $h_0(t)$. The results easily can be generalized to include other covariates and confounders in the model.

Since the baseline hazard $h_0(t)$ is not completely specified, it is difficult to use the full likelihood, which is the joint distribution of the failure times to estimate **b**. Cox in his seminal paper (9) introduced an innovative approach to estimate **b** without involving $h_0(t)$. The idea is to factor the full likelihood into two functions: one depending on both **b** and $h_0(t)$ and the other only on the parameters **b**. In the estimation process, the first function is ignored and the second is maximized with respect to **b**. The second function, which is referred to as a partial likelihood function, is given by

$$\prod_{i=1}^{k} \left[\exp\left(\sum_{j=1}^{p} b_j x_{ji} \right) \Big/ \sum_{l \in R(t_{(i)})} \exp\left(\sum_{j=1}^{p} b_j x_{ji} \right) \right],$$

where $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ are the k exact failure times and $R(t_{(i)})$ consists of all individuals whose survival times are at least $t_{(i)}$. This procedure is known as the maximum partial likelihood method. Observe that the censored observations do not contribute to the numerator but are used in the denominator. That is, the likelihood depends only on the rank of the failure times and not on the actual values.

It is well documented in the literature that the distribution of the $\hat{b}$ is asymptotically normal, with mean estimated by the maximum partial likelihood estimate

of **b** and the variance-covariance matrix estimated by the second derivative of the likelihood function. The loss of efficiency, i.e. the variance being larger than the variance obtained by using the complete likelihood, is small. This is remarkable in the sense that this holds regardless of the distribution of the baseline hazard. Standard errors of the estimates of **b** can be obtained and confidence intervals for **b** can be constructed. The null hypothesis that $\mathbf{b} = 0$ can be tested by calculating the value of the ratio between the estimate of **b** divided by its standard error. The resulting statistic is referred to as the Wald statistics and is compared to a tabulated value obtained from the standard normal distribution.

Tied observations, i.e. two or more patients having the same survival time, do occur in practice, mostly as a result of rounding. Several methods have been proposed to handle this situation. The simplest and most commonly used approach is due to Breslow (4). The partial likelihood function is approximated by

$$\prod_{i=1}^{k} \left\{ \exp\left(\sum_{j=1}^{p} b_j x_{ji} \right) \Big/ \sum_{1 \in R(t_{(i)})} \left[\exp\left(\sum_{j=1}^{p} b_j x_{j_i} \right) \right]^{m_{(i)}} \right.,$$

where $m_{(i)}$ denote the multiplicity of $t_{(i)}$. It is assumed that the $m_{(i)}$ deaths at $t_{(i)}$ are distinct and occurred sequentially. The approximation is poor when tied survival times are heavy. Alternative approximations include the works of Peto (40), Efron (16), and Cox (9). Cox's approach assumes that survival times are really discrete, as opposed to Breslow. He generalizes the model by a logistic transformation

$$\frac{h(t)dt}{1 - h(t)dt} = \frac{h_0(t)dt}{1 - h_0(t)dt} \exp\left(\sum_{j=1}^{p} b_j x_{ji} \right).$$

This model reduces to Equation 1 in the continuous case.

After **b** has been estimated, one may want to estimate the hazard function or equivalently, the survivorship function. Estimates of the baseline hazard function or its cumulative hazard are proposed by Kalbfleisch & Prentice (27) and Breslow (4). The properties of Breslow estimates are investigated by Tsiatis (59).

Similar to multiple regression problems, stepwise procedures provide an algorithm of ranking the covariates in order of their relative importance. From these rankings and by using significant tests on each variable, the most significant prognostic factors related to the dependent variable are selected.

How do we know if the assumption of proportional hazards holds? An easy way to check the hazards assumption is to plot the cumulative hazard functions. From Equation 1, it is easily shown that

$$\log_e[-\log_e S(t, x)] = -bx + \log_e[-\log_e S_0(t)],$$

where S(t, x) is the survival function. For two different sets of covariates x_1 and x_2, the functions $\log_e[-\log_e S(t, x_1)]$ and $\log_e[-\log_e S(t, x_2)]$ have a constant difference and hence, from the reference cumulative baseline hazard $\log_e[-\log_e S_0(t)]$, are parallel. The process is tedious with many covariates. One strategy is to stratify the data into several groups. For each group, compute the cumulative hazard function. Plot all the cumulative hazard functions in the same axes against time. If all the functions appear to be parallel, then there is no reason to believe that the proportional hazards assumption is not met. Formal tests are available to check this hypothesis. One straightforward technique is to use a time-dependent variable. Incorporate the time variable t or $\log_e t$ as an independent variable. To improve numerical stability, the time variable is centered about an arbitrary point t_0. The model is

$$h(t) = h_0(t) \exp[\mathbf{b}_1^T \mathbf{x} + b_2(\log_e t - \log_e t_0)].$$

Testing whether $b_2 = 0$ is equivalent to testing whether the proportional hazards assumption is tenable. Note that if $b_2 > 0$ ($b_2 < 0$), the hazard ratio increases (decreases) with time.

If the proportional hazards assumption does not hold, one remedy is to stratify the data into subgroups and apply the model for each stratum. Note that the hazards are nonproportional because the baseline hazards may be different between strata. The model is given by

$$h_{gi}(t) = h_{0g}(t) \exp(\mathbf{b}^T \mathbf{x}_{ig}).$$

The coefficients **b** are assumed to be the same for each stratum g. The partial likelihood function is simply the product of the partial likelihood functions in each stratum. A drawback of this approach is that the estimate of the effect of the stratifying variable can not be calculated.

Nonproportional hazards also occur when some covariates are time dependent. For example, status of cigarette smoking may change during the study period and therefore is time dependent. The setup of the partial likelihood functions of the time-independent covariates is still applicable. In this case, the covariates are indexed by time. The risk for survival is allowed to vary with time. The model takes into account the recent information about the covariates, though it does not model the changes that occurred in the variable over the observation period. This is in contrast to the time-independent model, which only includes baseline information.

A disadvantage should be mentioned. In comparing survival distributions, inclusion of time-dependent variables may confound the effect of treatment differences. For example, in comparing the time to recurrence of myocardial infarction between two treatment groups, if participants in one of the groups

improved significantly their health behaviors such as cessation of smoking and weight and blood pressure reduction, the true treatment effect may be confounded. However, the use of time-dependent variables still provides a better understanding of the relationship between the risk factors and survival.

In a study by Schwartzbaum et al (56) to determine the effect of prior exogenous estrogen use on survival after diagnosis of endometrial cancer, 244 newly diagnosed cases between 1970 and 1976 were followed to 1982. Of these, 46 were estrogen users and 198 were nonusers. Survival time was computed from the time of diagnosis. The proportional hazards assumption was checked by plotting the $\log_e[-\log_e \hat{S}(t)]$ for users and nonusers and for each stratum of potential confounding variables and effect modifiers. The resulting survival curves were approximately parallel. The authors concluded that there was no reason to reject the proportional hazards assumption.

The Cox regression was used to evaluate the effect of prior exogenous estrogen use for survival after diagnosis of endometrial cancer. Backward selection procedure was employed to determine the model. The authors started with the basic model containing the covariates: estrogen use (yes or no), age in years, histologic grade, and stage. The latter two are categorical variables. Potential effect modifiers (the interaction terms of estrogen use with the following variables: age, year of diagnosis, and obesity) were added to this basic model and tested for their significance. None was found to be significant. The last step was to add potential confounders to the basic model. These potential confounders included diabetes status, hypertension, obesity, race, year of diagnosis (1970–73 or 1974–76), and treatment type (radiation only, surgery only, radiation and surgery, other). Diabetes and hypertension were not significant and were dropped from the model. The results of the analyses are shown in Table 1.

The hazard ratios or relative risks were the ratios of the hazard rates for the values of a variable "unfavorable" to survival to the values of that variable "favorable" to survival. The favorable values are estrogen use; grade I; stages I and IA, except when compared with stage IB (then stage IB is favorable); obese; white race; surgery only, radiation only, and radiation and surgery; and year of diagnosis 1974–76. The unfavorable values are no estrogen use; grades II and III; stages II, III, and IV, and stages I and IA, only when compared with stage IB; nonobese; black race; treatment other than surgery only, radiation only, and radiation and surgery; and years of diagnosis 1970–73. For the continuous variable age at diagnosis, 55 (the first quartile) was used as favorable and 68 (the third quartile) unfavorable.

The adjusted hazard ratio was obtained by exponentiating the regression coefficient. The adjusted hazard ratio for estrogen use is 4.01 [exp(1.3889)].

Table 1 Results of Cox regression analysis on survival after diagnosis of diagnosis of endometrial cancer

Variable	Coefficient	Standard error	Hazard ratio	95% CI
Estrogen use	−1.3889	0.6082	4.01	1.22–13.21
Age at diagnosis	0.0290	0.0126	1.46	1.06–2.01
Grade				
II	0.7827	0.3146	2.19	1.18–4.05
III	0.6890	0.3655	1.99	0.97–4.08
Stage				
IB	−0.7897	0.4093	2.20	0.99–4.91
II	0.2677	0.3113	1.31	0.71–2.41
III, IV	1.2231	0.3385	3.40	1.75–6.60
Obesity	−0.6881	0.2537	2.00	1.21–3.27
Race	−0.8048	0.2874	2.24	1.27–3.92
Treatment				
Surgery only	−1.2910	0.5544	3.64	1.23–10.78
Radiation only	−0.4516	0.3243	1.57	0.83–2.97
Radiation and surgery	−1.6610	0.3687	5.26	2.56–10.84
Years of diagnosis	−0.6043	0.2744	1.83	1.07–3.13

Source: Schwartzbaum et al. 1987. The influence of exogenous estrogen use on survival after diagnosis of endometrial cancer. *Am. J. Epidemiol.* 126:851–60. Reproduced with permission from the *American Journal of Epidemiology*.

That is, the adjusted hazard ratio for nonusers is four times that of users. This showed a higher adjusted survival rate for estrogen users from endometrial cancer.

Another study that used the proportional hazards model, but with time-dependent variables, was a controlled clinical trial (6, 7, 53) on 415 patients with liver cirrhosis to evaluate the benefit of prednisone therapy. Two treatments were compared: prednisone and placebo. Of these 415 patients, 211 received prednisone and 204 were on placebo. The treatments were allocated randomly based on the date of birth. All patients underwent detailed assessment once a year and were observed up to 12 years.

The variables were divided into two types: prognostic and therapeutic. Prognostic variables are correlated with survival regardless of the treatment; whereas, for therapeutic variables, their correlation with survival depends on the type of treatment. Initially, a total of 162 variables were screened for inclusion in the analysis by logrank analysis. The strategy employed to determine whether a variable is prognostic or therapeutic is to fit separate models with the same set of covariates to both treatment groups. A covariate is prognostic if the two regression coefficients from each model could be replaced by a common coefficient and this coefficient is significantly different from zero. If the two coefficients are different and the difference between the two regression coefficients is significant then the covariate is considered to be therapeutic.

After the variables were classified as prognostic or therepeutic, the following model was fitted the data.

$$\mathrm{PI}(t)_{\mathrm{T}} = \log_{\mathrm{e}} \mathrm{h}(t)/\mathrm{h}_0(t) = \mathrm{b}_{\mathrm{tr}}\mathrm{z}_{\mathrm{tr}} + \mathrm{b}_1^{\mathrm{T}}\mathrm{z}_1(t) + \ldots + \mathrm{b}_{\mathrm{k}}^{\mathrm{T}}\mathrm{z}_{\mathrm{k}}(t) + \mathrm{b}_{\mathrm{k}+1}\mathrm{z}_{\mathrm{k}+1}(t) + \ldots + \mathrm{b}_{\mathrm{k}+\mathrm{r}}\mathrm{z}_{\mathrm{k}+\mathrm{r}}(t),$$

$\mathrm{PI}(t)_{\mathrm{T}}$ is referred to as the prognostic index at time t during treatment T. Higher values of $\mathrm{PI}(t)_{\mathrm{T}}$ would indicate higher risk or shorter survival. The first covariate, z_{tr}, is a treatment indicator that is time independent since each patient received only one treatment during the duration of the study. The therapeutic variables are given by $\mathrm{z}_1(t), \ldots,$ and $\mathrm{z}_{\mathrm{k}}(t)$ whose coefficients are indexed by superscript T. The value of the coefficient depends on the treatment type given to the patient. The prognostic variables are denoted by $\mathrm{z}_{\mathrm{k}+1}(t), \ldots, \mathrm{z}_{\mathrm{k}+\mathrm{r}}(t)$. Each of the variables is indexed by t to show that each may vary with time. These time-dependent variables are step functions, i.e. they are constant between annual examinations. At any time t, the values of these variables would be the values at the last follow-up immediately preceding t.

A number that measures the therapeutic effect of prednisone at time t is given by $\mathrm{PI}(t)_{\mathrm{placebo}} - \mathrm{PI}(t)_{\mathrm{prednisone}}$. This number, which was referred to as the therapeutic index, could be used as a guideline in determining whether to prescribe prednisone to new patients with cirrhosis. A higher value for this index indicates a beneficial effect of prednisone on survival. The statistical significance could also be tested by dividing the therapeutic index by its standard error. This statistic could then be compared to the tabulated values of the standard normal deviates.

The results from the stepwise procedure are shown in Table 2.

Treatment and age (at randomization) were time-independent variables. The other covariates are time dependent. The therapeutic variables are given by prothrombin index and two indicator variables for ascites (slight and moderate or marked). Each covariate has separate coefficients for the two treatment groups. The other variables in the Table 2 are prognostic variables. They include two continuous variables (albumin and alkaline phosphatase), one ordinal variable (daily alcohol consumption), and the rest (GI bleeding, bilirubin, liver connective tissue inflammation, and nutritional status) are indicator variables. The results showed that poor prognosis is associated with a low prothrombin index, marked ascites, GI bleeding, high age, high daily alcohol consumption, high bilirubin, low albumin, slight liver connective tissue inflammation, poor nutritional status, and high alkaline phosphatase.

Consider the therapeutic variables. Prothrombin index and ascites showed significant interaction with treatment. The table indicates that high prothrombin index and absence of ascites were associated with a beneficial effect of

Table 2 Results from Cox stepwise procedure to evaluate the benefit of prednisone on survival for patients with cirrhosis

Variable	Scoring	Treatment group	Regression coefficient	Standard error	P-value
Age (years)	Value-60	Both	0.052	0.0085	<0.0001
Treatment	Prednisone = 0 Placebo = 1	Both	−0.13	0.19	0.5
Prothrombin index (% of normal)	Log(value)-4	Prednisone Placebo	−1.58 −0.83	0.22 0.19	<0.0001 <0.0001
Ascites, slight	Present = 1 Otherwise = 0	Prednisone Placebo	0.96 0.34	0.25 0.27	<0.0001 0.21
Ascites, moderate or marked	Present = 1 Otherwise = 0	Prednisone Placebo	1.66 1.17	0.25 0.23	<0.0001 <0.0001
GI bleeding, marked	Present = 1 Otherwise = 0	Both	1.41	0.18	<0.0001
Alcohol consumption, daily	None = 0 10–50 g = 3 >50 g = 9	Both	0.14	0.024	<0.0001
Bilirubin (μmol/l)	<70 = 0 > = 70 = 1	Both	0.94	0.18	<0.0001
Albumin (g/l)	Log(value*10)-4	Both	−1.20	0.27	<0.0001
Liver connective tissue inflammation	None or slight = 0 Moderate or marked = 1	Both	−0.56	0.14	<0.0001
Nutritional status	Meager or cachetic = 1 Otherwise = 0	Both	0.56	0.16	<0.001
Alkaline phosphatase (KA units)	Log(value*10)-4	Both	0.36	0.11	<0.001

Reprinted from Christensen et al, 1986. Updating prognosis and therapeutic effect evaluation in cirrhosis with Cox's multiple regression model for time-dependent variables. *Scand J. Gastroenterol*., 21:163–74, by permission of Scandinavian University Press.

prednisone. Low prothrombin index and presence of ascites were associated with a harmful effect of prednisone.

To see how to use the prognostic and therapeutic indices in evaluating survival, consider a patient on prednisone with a prothrombin index of 60% of normal, no ascites, no GI bleeding, age 70 years, daily alcohol consumption of greater than 50 grams, bilirubin exceeding 70 μmol/l, albumin of 60 g/l, moderate liver connective tissue inflammation, poor nutrition, and alkaline phosphate of 8 KA; the prognostic index is -0.51. This corresponds to a risk of 60% ($e^{-0.51}$) compared to the baseline. For patient on placebo with the same set of covariates the index is given by -0.23. Hence, the difference between the two therapeutic indices is $+0.28$, which suggests a better chance of survival for the patient on prednisone.

PARAMETRIC APPROACH

The widespread accessibility of personal computers and the rapid advancement of statistical computing have offered impetus to the development of sophisticated nonparametric and semiparametric procedures in analyzing survival data. Procedures that in the past would normally take long hours of computing can now be done by the computer in just a few seconds. The traditional method of using a parametric approach in reliability investigations has been abandoned by many biomedical researchers during the past three or four decades in favor of nonparametric procedures. Nonparametric methods require fewer assumptions than parametric methods. No distributional assumption is imposed on the survival times. In addition, there are many physical causes that led to the failure or death of an individual, and it is very difficult, if not impossible, to mathematically take into account all these factors into a specific functional form.

Despite these advantages and popularity of nonparametric and semiparametric procedures, parametric modeling offers useful alternatives. Parametric models offer alternatives to the Cox model when the proportional hazards assumption is in question. If the distributional assumption on the survival times is valid, then the resulting estimates of the parameters, e.g. the median, are more efficient in the sense of having smaller standard errors as compared to those in the nonparametric models. Interpretability of the results is easier. In addition, the full likelihood is utilized in the analysis, resulting in more precise statistical inferences.

Parametric Models

A number of theoretical distributions have been used to approximate survival data. Several of these models are reviewed here.

The exponential distribution serves as the basic model for survival time, which is a special case of several more complicated distributions such as the Weibull and gamma. Table 3 lists six commonly used distributions in survival analysis with their density, $f(t)$, survival function, $S(t)$, and hazard function, $h(t)$.

The exponential distribution is characterized by a constant hazard function, λ, independent of the age of the individual. There is no aging or wearing out, and failure or death is a random event independent of time. Since $\log_e S(t) = -\lambda t$, a linear function of t, it is easy to determine whether data follow an exponential distribution by plotting $\log_e S(t)$ against t. A linear configuration indicates that the data come from an exponential distribution and the slope of the straight line is an estimate of the hazard rate λ.

Applications of the exponential distribution as a suitable model for survival distribution include the areas of cancer (57, 60), bacterial carriage (12),

Table 3 Commonly used distribution in survival data analysis

Distribution	Parameter	Density and survival functions	Hazard function
Exponential	$\lambda > 0$	$f(t) = \lambda \exp(-\lambda t)$ $S(t) = \exp(-\lambda t)$	$h(t) = \lambda$
Weibull	$\lambda, \gamma > 0$	$f(t) = \lambda\gamma(\lambda t)^{\gamma-1} \exp(-\lambda t)^{\gamma}$ $S(t) = \exp(-\lambda t)^{\gamma}$	$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$
Gamma	$\lambda, \gamma > 0$	$f(t) = [\lambda/\Gamma(\gamma)](\lambda\gamma)^{\gamma-1} \exp(-\lambda t)$ $S(t) = \int_t^{\infty} f(x)dx$	$h(t) = f(t)/S(t)$
Gompertz	$\lambda, \gamma > 0$	$f(t) = \exp[(\lambda + \gamma t) - 1/\gamma(e^{\lambda+\gamma t} - e^{\lambda})]$ $S(t) = \exp[-e^{\lambda}/\gamma(e^{\gamma t} - 1)]$	$h(t) = \exp(\lambda + \gamma t)$
Lognormal	$\mu, \sigma > 0$	$G(y) = \frac{1}{\sqrt{2\pi}} \int_0^{y} \exp(-u^2/2)\, du$	$h(t) = \{1/t\sigma\sqrt{2\pi}$
	$a = \exp(-\mu)$	$f(t) = 1/t\sigma\sqrt{2\pi} \exp[-1/2\,\sigma^2(\log_e at)^2]$	$\times \exp[-1/2\sigma^2(\log_e at)^2]\}$
		$S(t) = 1 - G(\log_e at/\sigma)$	$\times \{1 - G(\log_e(at/\sigma))\}^{-1}$
Log-logistic	$\lambda, \gamma > 0$	$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}[1 + (\lambda t)^{\gamma}]^{-2}$ $S(t) = [1 + (\lambda t)^{\gamma}]^{-1}$	$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$ $[1 + (\lambda t)^{\gamma}]^{-1}$

sensory-evoked potentials (14), trends of smoking cessation (15), national infant and child mortality (50).

The Weibull distribution is characterized by two parameters, γ (shape parameter) and λ (scale parameter). When $\gamma = 1$, the hazard rate remains constant as time increases; this is the exponential case. The hazard rate increases when $\gamma > 1$ and decreases when $\gamma < 1$ as t increases. Thus, the Weibull distribution may be used to model the survival distribution of a population with increasing, decreasing, or constant risk, and therefore has a broader application. Graphical presentations of $\log_e S(t) = -(\lambda t)^{\gamma}$ would assist an investigator to determine whether the data come from a Weibull distribution. When $\gamma = 1, \log_e S(t)$ is a straight line with negative slope. When $\gamma < 1$, negative aging, $\log_e S(t)$ decreases very slowly from zero and then approaches a constant value. When $\gamma > 1$, positive aging, $\log_e S(t)$ decreases sharply from zero as t increases. The Weibull distribution has been used, among other applications, to investigate carcinogenesis experiments by Williams (62), to characterize radiation response by Scott & Hahn (55), to investigate the time between release and return to prison for criminal recidivists by Schmidt & Witte (54), to model human mortality distribution by Juckett & Rosenberg (26), to analyze trends of smoking cessation by Elketroussi & Fan (15), and to model the treatment-free incubation period of AIDS by Hendriks et al (21).

The gamma distribution is also characterized by two parameters, γ (shape parameter) and λ (scale parameter). When $0 < \gamma < 1$, there is negative aging and the hazard rate decreases monotonically from infinity to λ as time increases from zero to infinity. When $\gamma > 1$, there is positive aging and the hazard rate

increases monotonically from zero to λ as time increases from zero to infinity. When $\gamma = 1$, the hazard rate equals λ, a constant, as in the exponential case. The gamma distribution has been applied to hepatograms in normal adults and in patients with cirrhosis and obstructive jaundice (18), speech amplitude (39), platelet survival (3), genetic influences on the age-at-menarche (36), and treatment-free incubation time of AIDS (21).

The Gompertz distribution is also characterized by two parameters, λ and γ. When $\gamma < 0$ (>0), the hazard rate decreases (increases) from $\exp(\lambda)$, and when $\gamma = 0$, the hazard rate is constant, $\exp(\lambda)$. The distribution was compared to the Weibull distribution as a descriptor for human mortality distributions by Juckett & Rosenberg (26) and to the exponential distribution and other models in breast cancer growth by Spratt et al (57). It was also used by Riggs in a series of papers on adult mortality due to a variety of diseases such as Parkinson's disease, lung cancer, prostate cancer, stroke, and emphysema (43–48).

The lognormal distribution can be defined as the distribution of a variable whose logarithm follows the normal distribution. Consider the random variable T such that $\log_e T$ is normally distributed with mean μ and variance σ^2. Then T is lognormally distributed with parameters μ and σ^2. It should be noted that μ and σ^2 are not the mean and variance of the lognormal distribution. The hazard function increases initially to a maximum and then decreases, almost as soon as the median is passed, toward zero as time approaches infinity. Thus, the lognormal distribution describes a first increasing and then decreasing hazard rate. The lognormal distribution has been applied frequently in reliability and biomedical research, partially due to its close relationship with the normal distribution. For example, Horner (23) found that the distribution of age at onset of Alzheimer's disease can be approximated by the lognormal distribution. Recently, Larsen et al (29) used a lognormal model to relate human lung function (FEV1) decrease to O_3 exposure, and Ahmed et al (1) found that the lognormal distribution was appropriate in a study of human health risk due to consumption of chemically contaminated fishery products. They reported that the lognomal distribution provided good descriptions of the pattern of variation of contaminant concentrations among different species and geographic areas. The results offer an opportunity to reduce exposure through restricting harvested fishery products from certain areas and by excluding certain species.

Similar to the lognormal distribution, if the variable $\log_e T$ has a logistic distribution, the variable T follows the log-logistic distribution. The log-logistic distribution also has two parameters, λ and γ. When $\gamma < 1$, the hazard rate decreases from infinity toward 0, when $\gamma = 1$, it decreases from λ to 0, and when $\gamma > 1$, it increases from 0 to a maximum and then decreases toward 0. The special features of the hazard function of the log-logistic distribution provide a good alternative to the Weibull, lognormal and gamma distributions. It has

been used to model the time to the return to prison in criminology (54) and the time between smoking cessation and restarting (15).

After a model is selected for the survival data, the parameters can be estimated by the maximum likelihood method in most cases. The likelihood function is in general

$$L = \prod_{i=1}^{r} f(t) \prod_{i=1}^{n-r} S(t)$$

where $f(t)$ and $S(t)$ are, respectively, the density function and the survivorship function of the parametric distribution selected, and n and r are the total number of observations and the number of uncensored observations, respectively. The goodness-of-fit can be evaluated by statistical tests (22, 24, 28), graphical methods (30, 38), or by examining the residuals (10). Probability of surviving longer than a given time (not limited to the range of the observed time period) easily can be estimated from the survivorship function, $S(t)$. In the following, an example is reviewed.

Elketroussi & Fan (15) analyzed the time trends of smoking cessation using six mathematical survival models with data from the Multiple Risk Factor Intervention Trial (MRFIT). This longitudinal study was designed to investigate the effect of reduced cardiovascular risk factors, including smoking. For illustrative purposes, only the exponential, Weibull and log-logistic models are discussed here. The study identified 12,866 men aged 35–57 years as having above-average risk of coronary heart disease based on their blood pressure, serum cholesterol, and number of cigarettes smoked per day. The participants were evenly divided into two groups, one group received special intervention (SI) and the other received the usual medical care (UC). Of the 6428 SI participants, 4103 (63.8%) were smokers at the time of initial visit. After an initial intensive intervention, 1676 smokers quit smoking within the first four months. These participants who quit smoking were monitored every four months, and the number of recidivists was recorded at every follow-up time. The second column of Table 4 gives the number of recidivists in each follow-up interval.

The "survival time" here is the time from initial abstention from smoking to resumption. The distribution of the time from initial abstention to resumption is clearly skewed, with most relapses occurring during the first 4–8 months. Censored cases are those who did not restart smoking during the study period and the censored times are from the time they quit to the end of the study. Parameters of the exponential, Weibull, and log-logistic models were estimated using the maximum likelihood method. Table 4 also gives the predicted numbers of recidivists by the three models. The authors used several criteria to evaluate the goodness-of-fit: (*a*) the maximum difference between the observed proportion of survivors

Table 4 Observed and predicted number of recidivists

Follow-up interval (months)	Number of persons: Observed recidivists (MRFIT)	Predicted recidivists: Exponential	Log-logistic	Weibull
0–4	288	173	292	294
4–8	149	135	120	115
8–12	55	105	78	77
12–16	46	82	57	57
16–20	47	64	44	45
20–24	32	50	35	36
24–28	24	39	29	30
28–32	32	30	25	25
32–36	32	24	21	21
36–40	14	19	18	18
40–44	17	14	16	16

Source: Elketroussi & Fan (15). Reproduced with permission from the *International Journal of Biomedical Computing.*

and the fitted value, (*b*) the maximum log-likelihood value, (*c*) the square-root of the sum of the squared error (RSSE) between the observed values and the model's projection, and (*d*) the RSSE between the projected proportion of survivors and the observed data plus/minus two standard deviations. Models for which the RSSE is within two standard deviations do not provide an adequate fit.

Among the three models, the Weibull and the log-logistic provided a better fit than the exponential. The maximum likelihood estimates (MLE) of γ and λ are, respectively, 0.618 and 0.058 per month. With $\gamma < 1$, it implied that the risk of recividism decreased over time. Similar results were obtained using the log-logistic model. The MLE of the γ and λ in this model were 0.743 and 0.068, again indicating a decreasing hazard function. The goodness-of-fit results based on the four criteria are given in Table 5. Every one of the four criteria indicates that the Weibull and the log-logistic models provided an

Table 5 Goodness-of-fit of the exponential, Weibull, and log-logistic models for the smoking cessation data

	Exponential	Weibull	Log-logistic
Max. Diff. in S(t)	4.5	1.6	1.4
Log (max. likelihood)	−2613	−2574	−2573
RSSE	7.5	2.1	2.0
RSSE ± 2 s.d.	0.13	0.0	0.0

Source: Elketroussi & Fan (15). Reproduced with permission from the *International Journal of Biomedical Computing*.

adequate fit to the data and the exponential fit was unsatisfactory. This is not unexpected since the data indicated that risk of restarting declined with time, not a constant as characterized by the exponential distribution.

Parametric Two-Sample Tests

If the survival distributions follow a known model, parametric tests of the parameters are more powerful than nonparametric. In the exponential case, the likelihood ratio test and an F-test suggested by Cox (8) can be used to test the equality of the two parameters, whether or not there are censored observations. If the survival time in the two groups follows the Weibull distribution, a two-sample test proposed by Thoman & Bain (58) can be applied, which tests the equality of the two-shape parameters. If the hypothesis $\gamma_1 = \gamma_1$ is rejected, we need not test the hypothesis $\lambda_1 = \lambda_2$, otherwise, we do need to test $\lambda_1 = \lambda_2$. To compare two gamma distributions, an F-test developed by Rao (42) is available. These parametric tests are also described in Lee (30).

The parametric two-sample tests are rarely used in public health research since finding an adequate model is not always easy and the nonparametric tests are readily available in most computer software packages and are equally good for practical purposes. Therefore, these tests are not reviewed here.

Parametric Regression Models

The Cox model is the most popular choice in survival data modeling. The facts that the underlying hazard function is arbitrary and that there is only a small loss in efficiency in using partial likelihood to estimate the parameters are very attractive. The model also encompasses several parametric models for the alternative function of $\log_e h_0(t)$ as defined in Equation 2. If $\log_e h_0(t)$ is a constant h, Equation 2 reduces to the exponential regression model. If $\log_e h_0(t) = ht$, Equation 2 gives the Gompertz, and if $\log_e h_0(t) = h \log_e t$, Equation 2 gives the Weibull regression model. The Weibull and the exponential regression models share the assumption of proportional hazards with the Cox regression model. Parametric models do offer useful alternatives. A plot of the empirical hazard function may suggest a family of parametric models. Estimates of the parameters in this case will be fully efficient and consistent. There are also situations where fitting a parametric model is justified, such as the violation of the proportional hazards assumption.

In parametric regression models, the dependence of survival time on covariates or risk factors is described explicitly through the parameters, hazard function, or survival function. For example, Breslow (4) uses the exponential distribution to model the remission duration of children with leukemia and to identify important prognostic factors. The parameter λ of the exponential distribution is assumed to be a function of the potential prognostic factors, for

example,

$$\lambda = \exp\left(\sum_{j=0}^{p} b_j x_j\right).$$

Estimation of the coefficients b_j's is carried out by using the maximum likelihood method and their significance tested by a standard normal test.

The accelerated failure time (AFT) model includes the exponential, Weibull, lognormal, gamma, and log-logistic distributions as special cases. This model is characterized by the following relationship between the survival functions of two individuals, *i* and *j*.

$S_i(t) = S_j(ct)$, where c is a constant. A most commonly used AFT model with covariates is to assume that

$$\log T = \sum_{j=0}^{p} b_j x_j + w\varepsilon,$$

where ε represents the random error, w is a scale parameter, and b_j's are coefficients to be estimated. The random error ε may follow any distribution such as the exponential, Weibull, lognormal, gamma, and log-logistic. Note that in the AFT model, the survival time cannot be zero.

Chapman et al (5) obtained data from 323 primary invasive node positive breast cancer patients, accrued at the Henrietta Banting Breast Center at Women's College Hospital, University of Toronto between January, 1977 and December, 1986. These patients were followed prospectively to 1990. Each patient completed a standard form containing medical and treatment histories and family history of cancer.

The objective of the study was to determine the important prognostic factors associated with recurrence of the cancer. The factors considered were (*a*) age (also dichotomized into ≤ 50 and >50 years, inferred menopausal status), (*b*) tumor size (cm, also categorized into T1, T2, and T3), (*c*) number of positive nodes (also categorized into 1–3, 4–6, 7–10, and ≥ 11), (*d*) estrogen receptor [fmol/mg protein, also categorized into $<10(-)$ and $\geq 10(+)$], (*e*) progesterone receptor [fmol/mg protein, also categorized into $<10(-)$ and $\geq 10(+)$], (*f*) combined estrogen (ER) and progesterone (PgR) receptors (ER−/PgR−, ER+/PgR−, ER−/PgR+, ER+/PgR+), (*g*) weight (kg, also categorized into ≤ 65 and >65), (8) adjuvant treatment (coded 0 to 4 as shown in Table 6 below).

Kaplan Meier survival functions were compared across different strata for each factor by Wilcoxon tests. The analyses yielded significant results for tumor size, grouped number of positive nodes, grouped estrogen receptor, and combination of estrogen and progesterone receptors. In addition, plots of the log cumulative hazards against time were examined across the same set of strata.

Table 6 Adjuvant treatment

AGE ≤ 50	AGE > 50	CODE
Chemotherapy + ovarian ablation $\pm$ hormones	Hormones	4
Chemotherapy + hormones	Chemotherapy + ovarian ablation $\pm$ hormones	3
Chemotherapy	Chemotherapy + hormones	2
Hormones	Chemotherapy	1
Nothing	Nothing	0

Source: Chapman et al (5). Reproduced with permission from Kluwer Academic Publishers.

The plots corresponding to grouped number of positive nodes and grouped adjuvant treatment suggested departure from the proportional hazards assumption.

The strategy employed was to investigate and compare the Cox model with a family of accelerated failure time models that included the exponential, Weibull, log-logistic, and lognormal. The fact that the hazards functions might not be proportional suggested one of the accelerated failure time models. Stepwise and all-subset variable procedures were carried out for each model. The results were reproduced in Table 7 below.

All of the models except the log-logistic model included number of positive nodes, tumor size, adjuvant treatment, and progesterone receptors. In addition to these four variables, the log-logistic model included the combined estrogen and progesterone receptors as the fifth variable.

The likelihood ratio (LR) tests were utilized in the comparison between models. Results of the comparison of the lognormal and with each of the other models (Cox, exponential, Weibull, log-logistic) supported the lognormal model. It was shown in other studies that the risk of recurrence of the disease increased to some point in time and then decreased. The same characteristic is shared by the hazard function of the lognormal. The final model was given by the lognormal model with standardized estimates given in Table 8 below. All

Table 7 Comparison of models

MODEL	-2 LOG LR
Cox	76.75
Exponential	82.32
Weibull	82.39
Log-logistic	91.19
Lognormal	86.81

Source: Chapman et al (5). Reproduced with permission from Kluwer Academic Publishers.

Table 8 Estimates for the lognormal model

VARIABLE	β/standard error	P-value
Number of positive nodes	−5.62	<0.01
Adjuvant treatment	3.71	<0.01
Tumor size	−4.01	<0.01
Progesterone receptors	2.32	<0.05

Source: Chapman et al (5). Reproduced with permission from Kluwer Academic Publishers.

variables were continuous. Higher number of positive lymph nodes and tumor size indicated shorter time and hence, higher risk to recurrence of the disease.

COMPUTER SOFTWARE FOR SURVIVAL DATA ANALYSIS

The analyses discussed above can easily be performed using commercially available computer software, which is brought about by the rapid development of statistical computing. Except for analyses that involve massive amounts of data, all computations can easily be carried out by microcomputers. The most popular software packages include SAS, BMDP, SPLUS, SPSS, EGRET, and GLIM, among others. The first two are discussed briefly in this section. Readers are advised to consult the manuals for the appropriate documentation of the different techniques discussed in this paper.

In SAS (2, 51, 52), survival analysis is covered in three major procedures: PROC LIFETEST, PROC LIFEREG, and PROC PHREG. PROC LIFETEST produces nonparametric estimates of the survival function using the lifetable method. For data with individual survival times, censored and uncensored, the Kaplan-Meier product limit estimates are generated. For grouped data, the actuarial method is used. The estimated survival curve and hazard function can be plotted in both linear and logarithmic scales, which allow an evaluation of whether any parametric models are appropriate for the survival time. In addition, it provides three methods, including the logrank test and Gehan's Wilcoxon test, for testing the equality of two or more survival distributions.

PROC LIFEREG fits parametric models, provides maximum likelihood estimates of the parameters, and evaluates the goodness-of-fit using graphical methods and likelihood ratio statistics. The procedure focuses on the AFT model, which includes the exponential, Weibull, lognormal, gamma, and log-logistic distributions. It can handle both categorical and numerical covariates; however, it cannot incorporate time-dependent covariates.

The procedure PROC PHREG implements Cox's proportional hazards model. It can handle continuous and discrete survival times and categorical and numeric

covariates including time-dependent covariates. If the assumption of proportional hazards does not hold, stratified analysis can be performed. In addition, estimates of the survival function with covariates incorporated are available.

In BMDP (13), there are two programs available for survival analysis, 1L and 2L. The program 1L has similar functions as PROC LIFETEST in SAS. It provides the Kaplan-Meier estimates of the survival function and lifetable analysis for grouped data. Graphical output include the survival curve, the hazard function, and the density function. The logrank test is referred to as the generalized Savage (Mantel-Cox) test. It also contains several other two-sample and K-sample tests that are not described in this paper. Stratified comparisons of two survival distributions are also available.

The program 2L has similar features as PROC LIFEREG and PROC PHREG in SAS, i.e. it provides estimates for Cox's proportional hazards model and the AFT model and plots. Stratified analysis is available to accommodate the case of nonproportional hazards. It allows time-dependent covariates and stepwise regression procedures in fitting Cox's model. The latter is not available in SAS. Similar to SAS, the following distributions are available in the AFT model in BMDP: exponential, Weibull, lognormal, and log-logistic. Stepwise procedures are also permitted in the AFT model. Program 2L also provides plots of residuals against the covariates in both the Cox (Cox-Snell residuals) and AFT (Cox-Snell and standardized residuals) models.

CONCLUDING REMARKS

Survival analysis has been widely used in public health research from medical and epidemiological investigations to health economics and social and behavioral studies. The relationship between a given event and potential "risk factors" or the estimation of a causal or predictive model is often of interest. Survival analysis is best suited for longitudinal studies where the event of interest is observed when it occurs, however, it is used frequently with retrospective data.

The methods reviewed in this paper assume that the survival times can be measured from a known and exact starting point of time and are either known exactly (uncensored) or are known to be longer than the observation time (censored). In reality, this may not be the case. For example, in studies of human immunodeficiency virus (HIV) disease, the time between infection of an individual and transmission of the virus to their sexual partners is often of great interest. Both events in this case are never exactly observed. We may know that the events occurred during certain periods of time and therefore the times of both events are "interval censored." The time period between the two events is never observed exactly. Such data are referred to as *doubly censored* by DeGruttola & Lagakos (11). This situation is recently revisited by Jewell et al (25).

The methods reviewed here also assume that the samples are independent. This assumption may also be violated in public health research. For example, in a mortality study of cardiovascular disease, the sample may include siblings or father and son and therefore observations are likely to be correlated. This situation has been discussed in several papers, e.g. Wei et al (61), Lee et al (31), Lipsitz et al (32), and most recently Lipsitz & Parzen (33).

Another frequently seen situation is recurrent events such as nonfatal myocardial infarction in patients with heart disease and repeated infections for kidney dialysis patients. The times between the recurrent episodes are not independent, and some patients are more prone or have an unknown tendency to have repeated events. This unknown tendency is referred to as *frailty*, which is considered to be a random variable with a distribution. The distribution is estimated by using the sample data including the covariates. McGilchrist & Aisbett (35) study the frailty of kidney dialysis patients.

In estimating the survival distribution, the Kaplan-Meier method has been the most commonly used. Recently, Murray & Tsiatis (37) propose a new method that uses prognostic covariate information to recover some of the information lost due to censoring. The estimate is more efficient than the Kaplan-Meier estimate when the covariate incorporated is prognostic, and it reduces to the Kaplan-Meier estimate when the covariate is not prognostic or there is no censoring.

Thus, survival analysis is a dynamic area in statistics, with many new methods in response to the practical needs in various applications. In most applications, nonparametric methods suffice. Thus, the Kaplan-Meier method, the logrank and Wilcoxon tests, and the proportional hazards model are most popular in survival data analysis. The parametric approach provides additional advantages. Although the selection of an adequate model is an art as well as a scientific task, recent development in computer software has made the application of parametric models much easier.

The Cox's proportional hazards model has been routinely applied in various investigations. It does not require a specific parametric hazard function and the incorporation of time-dependent covariates becomes convenient when using a commercially available computer package such as SAS or BMDP. The assumption of proportional hazards is often appropriate, and stratified analysis is helpful if the assumption is not met. However, there is still room for other models such as the accelerated failure time model family. More investigators (e.g. 5) are using parametric models and find the alternatives more appropriate in identifying prognostic or risk factors. Again, this may be partially attributable to the availability of software for the parametric models. With convenience, public health researchers should be encouraged to use different models to sort out the data and determine the most important risk or prognostic factors with more confidence.

Literature Cited

1. Ahmed FE, Hattis D, Wolke RE, Steinman D. 1993. Human health risks due to consumption of chemically contaminated fishery products. *Environ. Health Perspect. Suppl.* 101:297–302
2. Allison PD. 1995. *Survival Analysis Using The Sas System: A Practical Guide.* NC: SAS Institute Inc.
3. Bolin RB, Greene Jr. 1986. Stored platelet survival data analysis by a gamma model. *Transfusion* 26:28–30
4. Breslow N. 1974. Covariance analysis of censored survival data. *Biometrics* 30:89–99
5. Chapman JW, Trudeau ME, Pritchard KI, Sawka CA, Mobbs BG, et al. 1992. A comparison of all-subset Cox and accelerated failure time models with Cox step-wise regression for node-positive breast cancer. *Breast Cancer Res. Treat.* 22:263–72
6. Christensen E, Schlichting P, Andersen PK, Fauerholdt L, Juhl E, et al. 1985. A therapeutic index that predicts the individual effects of prednisone in patients with cirrhosis. *Gastroenterology* 88:156–65
7. Christensen E, Schlichting P, Andersen PK, Fauerholdt L, Shou G, et al. 1986. Updating prognosis and therapeutic effect evaluation in cirrhosis with Cox's multiple regression model for time-dependent variables. *Scand. J. Gastroenterol.* 21:163–74
8. Cox DR. 1953. Some simple tests for Poisson variates. *Biometrika* 40:354–60
9. Cox DR. 1972. Regression models and life tables. *J. R. Statist. Soc.* B34:187–220
10. Cox DR, Snell EJ. 1968. A general definition of residuals. *J. R. Statist. Soc.* B30:248–75
11. DeGruttola V, Lagakos SW. 1989. Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* 45:1–11
12. Dewals P, Bouckaert A. 1985. Methods for estimating the duration of bacterial carriage. *Int. J. Epidemiol.* 14:628–34
13. Dixon WJ, Brown MB, Engelman L, Jennrich RI. 1988. *BMDP Statistical Software Manual*, Vol 2. Los Angeles: Univ. Calif. Press
14. Eggermont JJ. 1988. On the rate of maturation of sensory evoked potentials. *Electroencephalogr. Clin. Neurophysiol.* 70: 293–305
15. Elketroussi M, Fan DP. 1991. Time trends of smoking cessation analyzed with six mathematical survival models. *Int. J. Biomed. Comput.* 27:231–44
16. Efron B. 1977. The efficiency of Cox's likelihood function for censored data. *J. Am. Statist. Assoc.* 76:312–19
17. Fenn P, McGuire A, Phillips V, Backhouse M, Jones D. 1995. The analysis of censored-treatment cost data in economic evaluation. *Med. Care* 33:851–63
18. Galli G, Maini CL, Salvatori M, Andreasi F. 1983. A practical approach to the hepatobiliary kinetics of $^{99M}T_c$-HIDA: clinical validation of the method and a preliminary report on its use for parametric imaging. *Eur. J. Nucl. Med.* 8:292–98
19. Gehan EA. 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 52:203–23
20. Gehan EA. 1965. A generalized two-sample Wilcoxon test for doubly-censored data. *Biometrika* 52:650–53
21. Hendriks JCM, Medley GF, Van Griensven GJP, Coutinho RA, Heisterkamp SH, et al. 1993. The treatment-free incubation period of AIDS in a cohort of homosexual men. *AIDS* 7:231–39
22. Hollander M, Proschan F. 1979. Testing to determine the underlying distribution using randomly censored data. *Biometrics* 35:393–401
23. Horner RD. 1987. Age at onset of Alzheimer's disease: clue to the relative importance of etiologic factors? *Am. J. Epidemiol.* 126:409–14
24. Hyde J. 1977. Testing survival under right censoring and left truncation. *Biometrika* 64:225–30
25. Jewell NP, Malani HM, Vittinghoff E. 1994. Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS. *J. Am. Statist. Assoc.* 89:7–18
26. Juckett DA, Rosenberg B. 1993. Comparison of the Gompertz and Weibull functions as descriptors for human mortality distributions and their intersections. *Mech. Ageing Dev.* 69:1–31
27. Kalbfleisch JD, Prentice RL. 1980. *The Statistical Analysis of Failure Time Data.* New York: Wiley

28. Koziol J, Green S. 1976. A Cramer-Von Mises statistics from randomly censored data. *Biometrika* 63:465–74
29. Larsen RI, McDonnell WF, Horstman DH. 1991. An air quality data analysis system for interrelating effects, standards, and needed source reducations: Part II. A log-normal model relating human lung function decrease to O_3 exposure. *J. Air Waste Manage. Assoc.* 41:455–59
30. Lee ET. 1992. *Statistical Methods for Survival Data Analysis.* New York: Wiley. 2nd ed.
31. Lee EW, Wei LJ, Amato DA. 1992. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, ed. JP Klein, PK Goel, 237–47. Boston: Kluwer
32. Lipsitz SR, Dear KBG, Zhao L. 1994. Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics* 50:842–46
33. Lipsitz SR, Parzen M. 1996. A jackknife estimator of variance for Cox regression for correlated survival data. *Biometrics* 52:291–98
34. Mantel N. 1967. Ranking procedures for arbitrarily restricted observations. *Biometrics* 23:65–78
35. McGilchrist CA, Aisbett CW. 1991. Regression with frailty in survival analysis. *Biometrics* 461–66
36. Meyer JM, Eaves LJ, Heath AC, Martin NG. 1991. Estimating genetic influences on age-at-menarche: a survival analysis approach. *Am. J. Med. Genet.* 39:148–54
37. Murray S, Tsiatis AA. 1996. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 52:137–51
38. Nelson W. 1972. Theory and applications of hazard plotting for censored failure data. *Technometrics* 14:945–66
39. Niederjohn RJ, Haworth DJ. 1986. Speech amplitude statistics and the gamma density function as a theoretical approximation. *J. Acoust. Soc. Am.* 80:1583–88
40. Peto R. 1972. Contribution to the discussion of paper by D.R. Cox. *J. R. Statist. Soc.* B34:205–7
41. Peto R, Peto J. 1972. Asymptotically efficient rank invariant procedures. *J. R. Statist. Soc.* A135:185–207
42. Rao CR. 1952. *Advanced Statistical Methods in Biometric Research.* New York: Wiley
43. Riggs JE. 1990a. Longitudinal Gompertzian analysis of adult mortality in the US, 1900–1986. *Mech. Ageing Dev.* 54:235–47
44. Riggs JE. 1990b. Longitudinal Gompertzian analysis of Parkinson's disease mortality in the US, 1955–1986: the dramatic increase in overall mortality since 1980 is the natural consequence of deterministic mortality dynamics. *Mech. Ageing Dev.* 55:221–33
45. Riggs JE. 1990c. Longitudinal Gompertzian analysis of stroke mortality in the US, 1951–1986: declining stroke mortality is the natural consequence of competitive deterministic mortality dynamics. *Mech. Ageing Dev.* 55:235–43
46. Riggs JE. 1991. Longitudinal Gompertzian analysis of lung cancer mortality in the US, 1968–1987, rising lung cancer mortality is the natural consequence of competitive deterministic mortality dynamics. *Mech. Ageing Dev.* 59:79–93
47. Riggs JE. 1991. Longitudinal Gompertzian analysis of prostate cancer mortality in the US, 1962–1987: a method of demonstrating relative environmental, genetic and competitive influences upon mortality. *Mech Ageing Dev.* 60:243–53
48. Riggs JE. 1992. The dynamics of aging and mortality in the United States, 1900–1988. *Mech Ageing Dev.* 66:45–57
49. Saeki S, Ogata H, Okubo T, Takahashi K, Hoshuyama T. 1995. Return to work after stroke: a follow-up study. *Stroke* 26:399–401
50. Sankrithi U, Emanuel I, Van Belle G. 1991. Comparison of linear and exponential multivariate models for explaining national infant and child mortality. *Int. J. Epidemiol.* 20:565–70
51. SAS Institute Inc. 1990. *SAS/STAT User's Guide*, Vols. 1, 2, Ver 6. NC: SAS Institute Inc. 4th ed.
52. SAS Institute Inc. 1991. *Technical Report P-217, SAS/STAT Software: The PHREG Procedure, Ver 6.* NC: SAS Institute Inc.
53. Schlichting P, Christensen E, Andersen PK, Fauerholdt L, Juhl E, et al. 1983. Prognostic factors in cirrhosis identified by Cox's regression model. *Hepatology* 3:889–95
54. Schmidt P, Witte DA. 1988. *Predicting Recidivism Using Survival Models*. NY: Springer-Verlag
55. Scott BR, Hahn FF. 1980. A model that leads to the Weibull distribution function to characterize early radiation response probabilities. *Health Phys* 39:521–30
56. Schwartzbaum JA, Hulka BS, Fowler WC Jr, Kaufman DG, Hoberman D. 1987. The influence of exogenous estrogen use on sur-

vival after diagnosis of endometrial cancer. *Am. J. Epidemiol.* 126:851–60
57. Spratt JA, Von Fournier D, Spratt JS, Weber EE. 1992. Decelerating growth and human breast cancer. *Cancer* 71:2013–19
58. Thoman DR, Bain LJ. 1969. Two sample tests in the Weibull distribution. *Technometrics* 11:805–15
59. Tsiatis AA. 1981. A large sample study of the estimate for the integrated hazard function in Cox's regression model for survival data. *Ann. Statist.* 9:93–108
60. Walle AJ, Al-Katib A, Wong GY, Jhanwar SC, Chaganti RSK, et al. 1987. Multiparameter characterization of L3 leukemia cell populations. *Leuk. Res* 11:73–83
61. Wei LJ, Lin DY, Weissfeld L. 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Statist. Assoc.* 84:1065–73
62. Williams JS. 1978. Efficient analysis of Weibull survival data from experiments on heterogeneous patient populations. *Biometrics* 34:209–22