

Survival Analysis in Clinical Trials: Past Developments and Future Directions



Thomas R. Fleming; D. Y. Lin

Biometrics, Vol. 56, No. 4. (Dec., 2000), pp. 971-983.

Stable URL:

<http://links.jstor.org/sici?sici=0006-341X%28200012%2956%3A4%3C971%3ASAICTP%3E2.0.CO%3B2-E>

Biometrics is currently published by International Biometric Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/ibs.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

EDITORS' INVITED PAPER

Survival Analysis in Clinical Trials: Past Developments and Future Directions

Thomas R. Fleming* and D. Y. Lin

Department of Biostatistics, Box 357232, University of Washington,
Seattle, Washington 98195, U.S.A.

*email: fleming@biostat.washington.edu

SUMMARY. The field of survival analysis emerged in the 20th century and experienced tremendous growth during the latter half of the century. The developments in this field that have had the most profound impact on clinical trials are the Kaplan–Meier (1958, *Journal of the American Statistical Association* **53**, 457–481) method for estimating the survival function, the log-rank statistic (Mantel, 1966, *Cancer Chemotherapy Report* **50**, 163–170) for comparing two survival distributions, and the Cox (1972, *Journal of the Royal Statistical Society, Series B* **34**, 187–220) proportional hazards model for quantifying the effects of covariates on the survival time. The counting-process martingale theory pioneered by Aalen (1975, *Statistical Inference for a Family of Counting Processes*, Ph.D. dissertation, University of California, Berkeley) provides a unified framework for studying the small- and large-sample properties of survival analysis statistics. Significant progress has been achieved and further developments are expected in many other areas, including the accelerated failure time model, multivariate failure time data, interval-censored data, dependent censoring, dynamic treatment regimes and causal inference, joint modeling of failure time and longitudinal data, and Bayesian methods.

KEY WORDS: Accelerated failure time model; Censoring; Competing risks; Counting process; Cox regression; Failure time; Hazard function; Kaplan–Meier estimator; Log-rank statistic; Martingale; Multivariate failure times; Partial likelihood; Proportional hazards; Sequential analysis; Survival data.

1. Introduction

Enormous progress has been achieved in the development of the science of clinical trials during the 20th century. In this progress, methods have been developed, implemented, and refined that enable the reliable, efficient, and ethical evaluation of the benefits and risks of interventions that target the treatment and prevention of human diseases. One of the most important components of this development has been the formulation of censored data survival analysis methods.

The primary outcome measure in a clinical trial designed to provide a reliable assessment of benefit and risk often is defined to be the time to occurrence of a clinically important event, such as death, detection or progression of a disease, or occurrence of a clinically significant morbid event such as a serious infection, stroke, or major organ failure. A complexity that frequently arises in trials having time-to-event endpoints is that a substantial fraction of the trial participants remain free of the study endpoint at the time of data analysis. The patients who provide this incomplete outcome information are referred to as being censored or, more precisely, right censored

since it is only known that the true time-to-event for that participant exceeds the duration of follow-up.

The complexities provided by the presence of censored observations led to the development of a new field of statistical methodology. Because the analysis of clinical trials data with time-to-death outcomes provided the original motivation for this new statistical methodology, the field has become known as survival analysis. The methodological developments in this field, largely achieved in the latter half of the 20th century, in turn have had an enormous impact on the science of clinical trials.

The field of survival analysis is very rich. Thus, it is impossible to provide a comprehensive review of all the important developments in a single article, let alone to mention all the researchers who have made significant contributions to the field. This article provides an overview of some of the developments in survival analysis that we recognize to have been particularly influential on the science of clinical trials. The technical material will be kept to a minimum so that the article will be accessible to readers not specialized in survival analysis.

2. Notation and Identifiability

Let T denote the true time-to-event or failure time for a study participant in a clinical trial. The participant's time 0, i.e., where $T = 0$, represents the time of initiation of their follow-up in the clinical trial and is typically the time of randomization in a comparative study. Primary interest usually lies in estimation and testing regarding the distribution of T . This distribution can be characterized by the survival function $S(t) \equiv \Pr(T > t)$ for $t > 0$. Because of censoring, it is more convenient to deal with the hazard function. If T is continuous with density function f , then the hazard function is defined by

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \Pr(t \leq T < t + \Delta t \mid T \geq t) / \Delta t = f(t) / S(t).$$

The function $\Lambda(t) \equiv \int_0^t \lambda(u) du$ is called the cumulative hazard function for T , and it is easily shown that $S(t) = e^{-\Lambda(t)}$ for a continuous survival time T .

Let U denote the censoring time, i.e., the time beyond which the clinical trial participant cannot be observed. Then (T, U) are referred to as latent data, while the observed data are denoted by (X, δ) , where $X = \min(T, U)$, $\delta = I(T \leq U)$, and $I(\cdot)$ is the indicator function. The clinical trial participants having $\delta = 0$ are referred to as having censored observations.

While the distribution function $S(t)$ can be consistently estimated when data are uncensored, Tsiatis (1975) and Peterson (1976) established that neither $\lambda(t)$ nor $S(t)$ is identifiable or consistently estimable if one only observes (X, δ) . As discussed by Fleming and Harrington (1991, Theorem 1.3.1), observing (X, δ) rather than T for all participants only allows one to consistently estimate $S^\#(t) \equiv \exp\{-\int_0^t \lambda^\#(u) du\}$ for all t such that $\Pr(X > t) > 0$, where

$$\lambda^\#(t) = \lim_{\Delta t \downarrow 0} \Pr(t \leq T < t + \Delta t \mid T \geq t, U \geq t) / \Delta t. \quad (2.1)$$

Chiang (1968) referred to $\lambda^\#(t)$ as the crude hazard and $\lambda(t)$ as the net hazard. Therefore, in most survival analysis applications, a key assumption is made regarding the equality of the crude hazard (that is estimable) and the net hazard (that is of interest), i.e.,

$$\lambda^\#(t) = \lambda(t) \quad \text{for all } t \text{ such that } \Pr(X > t) > 0. \quad (2.2)$$

A sufficient condition for the validity of assumption (2.2) is the independence of T and U . Cox (1959), Mann, Schafer, and Singpurwalla (1974), Gross and Clark (1975), and Cox and Oakes (1984) provided further discussion.

An understanding about the mechanisms causing censoring provides evidence about validity of assumption (2.2). For example, if censoring is due to staggered entry (i.e., participants are enrolled into the trial over a lengthy accrual period and then followed to a common calendar date of analysis), T and U will essentially be independent as long as characteristics of participants that influence risk of events do not vary systematically over the period of accrual. In contrast, if censoring occurs when patients become too ill to be readily followed or if follow-up is stopped when participants are no longer compliant with their study intervention, censorship can be highly dependent, likely resulting in $\lambda^\#(t) \ll \lambda(t)$. In such settings,

statistical methods discussed in the next section that consistently estimate $S^\#(t)$ would substantially overestimate the true $S(t)$.

3. Estimation of the Hazard and Survival Functions

The early efforts in development of survival analysis methodology were predominantly focused on estimation of the hazard function $\lambda(t)$ and the survival function $S(t)$. Some parametric models were proposed, with maximum likelihood methods used for estimation of parameters, under the assumption of independent censoring (Mann et al., 1974; Gross and Clark, 1975). For example, assuming a constant hazard function, i.e., $\lambda(t) = \lambda$ for all $t > 0$, one obtains the exponential distribution, where the maximum likelihood estimator for λ is the number of observed events divided by the summation of duration of follow-up over all participants (Bartholomew, 1957). Kalbfleisch and Prentice (1980) explored a four-parameter generalized F family that incorporates many well-known parametric distributions having increasing, decreasing, and non-monotonic hazard functions. Among these distributions are the exponential, Weibull, log-logistic, gamma, and log-normal distributions. Lawless (1982) provided a detailed presentation of parametric methods.

A class of actuarial estimators were proposed to provide more robust estimation of $S(t)$. Generally, piecewise parametric models were fit to the data, where typically either the hazard function was assumed to be piecewise constant or the survival function was assumed to be piecewise linear. One would then estimate S_k , the conditional probability of survival to the end of the k th interval, given survival to the beginning of that interval. Crowley (1970) provided an overview of these estimators and explored their properties. The most widely used actuarial estimator was obtained by estimating S_k with $D_k / (Y_k - W_k)$, where, in the k th interval, D_k is the number of observed events, Y_k is the number of participants at risk at the beginning, and W_k is the number of participants censored before an event. This was called the standard life table estimator by Crowley (1970) and was proposed and initially explored by Berkson and Gage (1950) and Cutler and Ederer (1958). The variance for these actuarial estimators was usually obtained using the formula given by Greenwood (1926).

Nelson (1969) introduced hazard plotting based on his estimator of the cumulative hazard function $\Lambda(t)$. In the actuarial estimation approach, as the mesh of the intervals approaches zero, Nelson's estimator is given by a step function, with steps occurring at times of observed events and having size D/Y , where D events occur among Y participants at risk. Later, Johansen (1983) derived the Nelson estimator as a generalized maximum likelihood estimator.

Recognizing the relationship between $S(t)$ and $\Lambda(t)$ through the differential equation $-\{dS(t)/dt\}/S(t-) = \lambda(t)$, one motivates the relationship

$$-\{\Delta \hat{S}(t)\} / \hat{S}(t-) = \Delta \hat{\Lambda}(t),$$

where one estimates $\Lambda(t)$ using Nelson's estimator and then recursively solves for the estimator of $S(t)$. The resulting estimator is that proposed by Kaplan and Meier (1958). It is a step function, with value reduced by the multiplicative factor $\{1 - (D/N)\}$ at times of observed events.

Crowley (1970) established that, as the mesh of the actuarial intervals converges to zero, the Kaplan-Meier esti-

mator can be obtained as the exact limit of the standard life table estimator and the limit of the order $o(1/Y_k)$ for the other actuarial estimators. Kaplan and Meier (1958) had derived their estimator through a nonparametric maximum likelihood approach. Aalen (1975) established the estimator to be uniformly strongly consistent for $S(t)$ under the assumption that T is absolutely continuous and independent of U . (It is strongly consistent for $S^\#(t)$ whether or not independence holds.) Breslow and Crowley (1974) established the weak convergence of a standardized Kaplan–Meier estimator process to a time-transformed Brownian motion process, and Gill (1983) provided further development of the large-sample properties. Efron (1967) considered a two-sample application of the Kaplan–Meier estimator that he recognized to have a self-consistency property. The optimality of the Kaplan–Meier estimator was established by Wellner (1982).

4. Counting Processes and Survival Analysis, with Application to the Nelson Estimator

In the mid-1970s, Aalen introduced an elegant martingale-based approach to survival analysis, where statistical methods can be cast within a unifying counting process framework (see Aalen, 1975). In this seminal work, the counting process approach uses an integral representation for censored data statistics that provides a simple unified form for estimators, test statistics, and regression methods. These martingale methods allow one to obtain simple expressions for moments of complicated statistics and asymptotic distributions for test statistics and estimators and to examine the operating characteristics of censored data regression methods. Detailed presentation of this approach has been provided in textbooks by Fleming and Harrington (1991) and Andersen et al. (1993).

In the counting process approach for analyzing data on time-to-a-single-event, the data for the i th participant, (X_i, δ_i) , is represented as $\{N_i(t), Y_i(t)\}$ ($t > 0$), where

$$N_i(t) = I(X_i \leq t, \delta_i = 1) \quad \text{and} \quad Y_i(t) = I(X_i \geq t). \quad (4.1)$$

The right-continuous process $N(t)$ is referred to as the counting process since it essentially counts the number of events observed up to and including time t , while the left-continuous process $Y(t)$ is referred to as the at-risk process, indicating whether the participant is at risk at time t .

A simple yet important illustration of the counting process approach is provided by examining the properties of the Nelson estimator $\hat{\Lambda}(t)$ of $\Lambda(t)$. The hazard integrated over the region in which one has data is

$$\Lambda^*(t) \equiv \int_0^t I\{\bar{Y}(u) > 0\} \lambda(u) du,$$

where $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$ and n is the sample size. One can write $\hat{\Lambda}(t) - \Lambda^*(t)$ in the form of

$$\sum_{i=1}^n \int_0^t H_i(u) dM_i(u), \quad (4.2)$$

where $H_i(t) \equiv I\{\bar{Y}(t) > 0\}/\bar{Y}(t)$ is a left-continuous process and

$$M_i(t) \equiv N_i(t) - \int_0^t Y_i(u) \lambda(u) du \quad (4.3)$$

is the participant-specific martingale (with respect to a proper filtration). The martingale M_i in (4.3) represents the difference over the interval $(0, t]$ between the observed number and the model-predicted number of events for the i th participant. The left-continuity (and hence predictableness) of the process H_i and the martingale property for M_i renders the entire expression in (4.2) to be a martingale transform. This structure directly yields moments and large-sample properties. For example, since the martingale M_i has expectation zero, it follows that the Nelson estimator $\hat{\Lambda}(t)$ has expectation $\int_0^t \Pr\{\bar{Y}(u) > 0\} \lambda(u) du$. This martingale-based approach enables an elegant development of the small- and large-sample properties of the Nelson and Kaplan–Meier estimators, as shown by Gill (1980).

5. Two-Sample Statistics

The primary objective of many clinical trials is to provide a reliable comparison of the efficacy and safety of two interventions, where efficacy often is assessed in terms of a time-to-event outcome measure. Therefore, two of the most important achievements in the development of clinical trials methodology have been the acceptance of the need for randomization and the formulation of two-sample methods for comparing survival distributions in censored data.

In the 1930s, Fisher (1935) advocated that individuals who were eligible to participate in a clinical trial comparing two regimens be assigned to these two groups in a random manner. This process would eliminate the systematic occurrence of imbalances between intervention groups in the characteristics of study participants. In turn, the resulting lack of confounding between interventions and the patient characteristics would enable an unbiased evaluation of efficacy and safety, as long as the outcome variables were assessed on all randomized participants.

Maintaining the integrity of randomization in comparative clinical trials having long-term clinical outcomes and only partial follow-up on study participants as well as providing efficient statistical methods motivated the development of two-sample censored data statistics. A variety of parametric and nonparametric two-sample statistics have been proposed to assess observed differences in empirical survival curves. Parametric methods were described by Lawless (1982) and Kalbfleisch and Prentice (1980).

Two-sample censored-data linear rank statistics have provided a robust alternative to parametric methods. The most widely used member of this class is the log-rank statistic. It was originally proposed by Mantel (1966), who classified participants at risk at the time of an event into a 2×2 table according to event status (yes versus no) and intervention group. He then obtained the numerator of the log-rank statistic by computing the observed and the expected (conditioning on the margins of the 2×2 table) events on the control arm and by summing the differences of these over all distinct event times. Within each 2×2 table, the variance of the number of events on the control arm was obtained using the hypergeometric distribution. These were then summed over all distinct event times to provide the variance estimator for the log-rank statistic.

Aalen (1978) recognized that a wide class of two-sample statistics, including the log-rank statistic, could be written as a martingale transform as formulated in equation (4.2)

through proper choice of the left-continuous process H_i and where the sum over i would be taken over all participants in the two samples. Using this structure, Gill (1980) established several key properties of the log-rank statistic, including that it has mean zero when survival distributions are equal, that Mantel's variance estimator is unbiased, and that the statistic is asymptotically normal.

Gill (1980) introduced and explored a wide class of censored data two-sample statistics, called statistics of the class K . These are also called weighted log-rank statistics since any member of this class can be written as a weighted sum of the observed minus expected events in Mantel's 2×2 tables. The weight function is assumed to be a left-continuous (or, more generally, a predictable) process.

Many censored data two-sample statistics fall within Gill's class K . Some of the well-known examples are the censored data generalizations of the Wilcoxon statistic proposed by Gehan (1965), Efron (1967), and Peto and Peto (1972), the one-parameter class of statistics proposed by Tarone and Ware (1977), the G^ρ family proposed by Harrington and Fleming (1982), and its generalization, the $G^{\rho,\gamma}$ family, proposed by Fleming and Harrington (1991).

The statistics of the class K can be formulated as in equation (4.2). Using this structure, Gill (1980) derived small- and large-sample properties for statistics of this wide class. He developed criteria for consistency of these tests against ordered hazards and stochastic ordering alternatives. He also provided asymptotic distribution results, not only under the null hypothesis of equality of survival distributions but also under contiguous alternatives, allowing him to provide a characterization of the alternatives against which tests of the class K are efficient. Among these results was a proof that the log-rank statistic provides an efficient test under proportional hazards alternatives.

The efficient score function can be used to construct efficient rank tests for uncensored data (Hájek and Sidák, 1967). Prentice (1978) adapted these statistics to the censored data setting. Generally, these statistics either are of the class K or are asymptotically equivalent to such statistics. The results by Gill (1980), as well as work by Mehrotra, Michalek, and Mihalko (1982), Cuzick (1985), and Struthers (1984), establish that Prentice's adapted rank tests remain efficient in the presence of censoring.

The martingale formulation in (4.2) allows for straightforward development of the joint distribution of several statistics of the class K as well as supremum versions of such statistics. Using this structure, the properties of procedures based on clusters or supremum versions of statistics of the class K have been explored by many authors, including Gill (1980), Fleming et al. (1980), Fleming and Harrington (1981), Schumacher (1984), and Fleming, Harrington, and O'Sullivan (1987).

The martingale formulation in (4.2) also enabled the development of properties of weighted Kaplan-Meier statistics (Pepe and Fleming, 1991). These statistics are based on the integrated weighted difference in Kaplan-Meier estimators. Being direct generalizations of the t -test to censored data, they are sensitive against stochastic ordering alternatives.

6. Regression Models

In clinical trials designed to assess the effect of an intervention on a time-to-event outcome, it is important to be able

to explore or adjust for the effect of an array of other covariates that may be associated with that outcome. Hence, the information collected on each study participant is expanded to be (N, Y, \mathbf{Z}) , where \mathbf{Z} represents covariates. The covariates are usually assessed at time of enrollment of a study participant. They can be demographic variables, such as age, gender, or race; laboratory measurements, such as levels of bilirubin, blood pressure, or viral load; histologic assessments based on biopsy; or other descriptive measurements such as time from diagnosis of disease, type of disease, prior therapeutic exposures, or functional status of the participant. In regression models, these covariates can take a variety of functional forms, being dichotomous, discrete, or continuous. The continuous variables may be transformations of original measures, such as the logarithm of bilirubin.

The linear regression model for survival time data takes the form

$$\log T = \beta' \mathbf{Z} + \epsilon, \quad (6.1)$$

where β is a set of unknown regression parameters and ϵ is an error variable independent of \mathbf{Z} . The logarithmic transformation is employed because T is positive; other appropriate transformations of T may also be selected. Exponentiation of (6.1) yields $T = e^{\beta' \mathbf{Z}} T_0$, where $T_0 = e^\epsilon$. This expression shows that the role of \mathbf{Z} is to accelerate (or decelerate) the time to failure. Thus, (6.1) is referred to as the accelerated failure time model.

Because of censoring, it is more convenient to model the survival data through the hazard function. Let $\lambda(t | \mathbf{Z})$ denote the hazard function associated with \mathbf{Z} , i.e.,

$$\lambda(t | \mathbf{Z}) = \lim_{\Delta t \downarrow 0} \Pr(t \leq T < t + \Delta t | T \geq t, \mathbf{Z}) / \Delta t.$$

The proportional hazards model specifies that

$$\lambda(t | \mathbf{Z}) = \lambda_0(t) e^{\beta' \mathbf{Z}}, \quad (6.2)$$

where $\lambda_0(t)$ is the so-called baseline hazard function, i.e., the hazard function under $\mathbf{Z} = \mathbf{0}$, and β is a set of unknown regression parameters. Under this model, the covariates have multiplicative effects on the hazard function, and the regression parameters are interpreted as the logarithms of the hazard ratios or relative risks. The simplest form of (6.2) in which $\lambda_0(t)$ is a constant was first studied by Feigl and Zelen (1965).

Model (6.1) can be rewritten as

$$\lambda(t | \mathbf{Z}) = \lambda_0 \left(t e^{-\beta' \mathbf{Z}} \right) e^{-\beta' \mathbf{Z}}, \quad (6.3)$$

where $\lambda_0(t)$ is the hazard function of T_0 . A comparison of (6.3) with (6.2) reveals that the only overlap in the accelerated failure time and proportional hazards models arises when $\lambda_0(t)$ is Weibull (Kalbfleisch and Prentice, 1980, pp. 34-35).

In the regression setting, the independent censoring assumption given by equation (2.2) is extended so that, conditional on \mathbf{Z} , the crude and net hazard functions are equal. Survival models, such as (6.1) and (6.2), are referred to as parametric models if the distributional form of the failure time, i.e., $\lambda_0(t)$, is specified and as semiparametric models otherwise. Analysis of parametric survival models has been discussed by Kalbfleisch and Prentice (1980), Lawless (1982), Cox and Oakes (1984), and Andersen et al. (1993). Due to the complex nature of human diseases, it is difficult to specify the

parametric form. Thus, semiparametric models are preferable to parametric models in most clinical applications.

7. Cox Proportional Hazards Model

In his seminal papers in 1972 and 1975, Cox introduced an ingenious semiparametric approach to inference based on the proportional hazards model. These methodological results are among the developments in the field of survival analysis that have had the most profound impact on clinical trials applications.

By fitting the proportional hazards model in equation (6.2) with an unspecified baseline hazard function $\lambda_0(t)$, Cox obtained a robust approach for studying the influence of covariates on outcome. However, with an infinite-dimensional nuisance function $\lambda_0(t)$, modifications to the classical likelihood approach would be needed. Thus, Cox (1975) introduced the partial likelihood, which is based on the data that does not carry information about $\lambda_0(t)$. Specifically, Cox discarded the times of observed events and the number of events at those times. Assuming that censoring is independent and is uninformative for β (see Fleming and Harrington, 1991, Definition 4.3.1), he also discarded the censoring times and the identity of participants associated with the censored times. The partial likelihood then was based on, for all event times, the identity of the participant(s) failing at each event time, given the number failing and the identity of the participants at risk at that time. It takes the form

$$L(\beta) = \prod_{i \in \mathcal{D}} \frac{e^{\beta' \mathbf{z}_{(i)}}}{\sum_{j \in \mathcal{R}_i} e^{\beta' \mathbf{z}_j}}, \quad (7.1)$$

where \mathcal{D} is the set of indices of observed event times, $\mathbf{z}_{(i)}$ is the covariate vector for the subject failing at the i th observed event time T_i^0 , and \mathcal{R}_i is the set of participants at risk at T_i^0 . The maximum partial likelihood estimator $\hat{\beta}$ is the value of β that maximizes $L(\beta)$. Given $\hat{\beta}$, the cumulative baseline hazard function $\Lambda_0(t) \equiv \int_0^t \lambda_0(u) du$ is estimated by

$$\hat{\Lambda}_0(t) \equiv \sum_{i \in \mathcal{D}; T_i^0 \leq t} \frac{1}{\sum_{j \in \mathcal{R}_i} e^{\hat{\beta}' \mathbf{z}_j}}. \quad (7.2)$$

Estimator (7.2) is commonly attributed to Breslow (1972, 1974).

Cox (1972, 1975) conjectured that $L(\beta)$ shares the asymptotic properties of a full likelihood. A number of authors investigated the asymptotic properties of $\hat{\beta}$ and $\hat{\Lambda}_0(t)$. The first published proof was provided by Tsiatis (1981). Other proofs were given by Liu and Crowley (1978), Sen (1981), Naes (1982), and Bailey (1983). Andersen and Gill (1982) provided an elegant asymptotic theory for $\hat{\beta}$ and $\hat{\Lambda}_0(t)$ by observing that the partial likelihood score function can be formulated as a martingale transform of the form given in (4.2).

The relationship between partial likelihood inference in the proportional hazards model and a marginal likelihood approach was discussed by Kalbfleisch and Prentice (1973), while Jacobsen (1984) and Bailey (1979) discussed the relationship with a generalized maximum likelihood approach. The efficiency of $\hat{\beta}$ was studied by Efron (1977), Oakes (1977), and Cox and Oakes (1984).

One of the most important applications of the Cox proportional hazards model in the setting of a randomized controlled clinical trial is to obtain an adjustment of the estimator of treatment effect. Including a strongly predictive covariate in addition to the treatment variable in the regression model provides three potential benefits in the estimation of the regression coefficient of the treatment variable. The first is to address confounding, should imbalances arise between the covariate and the treatment group. The second is to address conservatism if the covariate were used as a stratification factor at randomization since failure to include such a covariate in the regression model would lead to overestimation of variance and conservative test procedures (see Anderson, 1989). The third is to address attenuation since biased estimators of the treatment effect would arise if the covariate were not included in the model (see Gail, Wieand, and Piantodosi, 1984; Lagakos and Schoenfeld, 1984; Anderson and Fleming, 1995; Fleming and Harrington, 1991, Example 4.2.1.).

8. Multiplicative Intensity Model

In many clinical trials, the outcome of primary interest extends beyond the time of the first event to exploration of the rate of recurrent events over time. These recurrent events may be repeated otitis media infections in an infant or repeated hospitalizations in an adult with a serious disease. To analyze such data, Aalen (1978) introduced the multiplicative intensity model as a generalization of the proportional hazards model. In this model, the participant-specific martingale is

$$M(t) = N(t) - \int_0^t Y(u) \lambda_0(u) e^{\beta' \mathbf{z}(u)} du, \quad (8.1)$$

where N and Y are of more general forms than given in (4.1). Specifically, the right-continuous counting process $N(t)$ still reflects the number of events that have occurred by time t but now has a range over all nonnegative integers. The at-risk process $Y(t)$ can be any left-continuous (or more broadly, predictable) process indicating, by one versus zero, whether or not the participant is at risk at time t . In addition, the covariate vector is allowed to be a (predictable) process. Self and Prentice (1982) discussed the link between the proportional hazards model and the multiplicative intensity model and the subtleties associated with time-dependent covariates.

In the semiparametric setting where $\lambda_0(t)$ in (8.1) is unspecified, one can use the partial likelihood principle to make inference about β and the Breslow estimator to estimate $\Lambda_0(t)$, although now the set \mathcal{D} in (7.1) and (7.2) may involve multiple event times from the same participant. The corresponding large-sample theory was again provided by Andersen and Gill (1982).

9. Regression Model Diagnostics

Extensive development of residuals has provided a wide variety of model diagnostics that are useful for the Cox proportional hazards model as well as for the broader multiplicative intensity model. Grambsch, Therneau, and Fleming (1998), in giving an overview of this development, summarize three classes of residuals. These are generalized residuals of Cox and Snell (1968), residuals based on counting process martingales and their transformations, and residuals from generalized linear regression models (McCullagh and Nelder, 1989) for log-linear Poisson regression.

These latter two classes of residuals can be briefly introduced. For simplicity, consider the participant-specific martingale in equation (4.3) for the special case of the Cox model given by (6.2). The corresponding martingale residual is

$$\hat{M}_i(t) \equiv N_i(t) - \hat{\Lambda}_0(t \wedge X_i) e^{\hat{\beta}' \mathbf{Z}_i},$$

where $a \wedge b = \min(a, b)$. This residual, introduced by Barlow and Prentice (1988) and explored by Therneau, Grambsch, and Fleming (1990), can be interpreted as the observed minus estimated model-predicted events for participant i over the interval $(0, t]$. As $t \rightarrow \infty$, the martingale residual reduces to

$$\hat{M}_i \equiv \delta_i - \hat{\Lambda}_0(X_i) e^{\hat{\beta}' \mathbf{Z}_i}.$$

The martingale residual \hat{M}_i is also within the class of Poisson residuals (Aitkin and Clayton, 1980; Whitehead, 1980; Laird and Olivier, 1981; McCullagh and Nelder, 1989). These residuals, symmetrized using the deviance transformation (McCullagh and Nelder, 1989), can be used to detect outliers. The partial residuals, defined by

$$\hat{M}_i / \left\{ \hat{\Lambda}_0(X_i) e^{\hat{\beta}' \mathbf{Z}_i} \right\} + \hat{\beta}_j Z_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, p,$$

where Z_{ij} and $\hat{\beta}_j$ are the j th components of \mathbf{Z}_i and $\hat{\beta}$, can be used to suggest the proper functional form for covariates in the model.

A class of martingale-transform residuals can be obtained by replacing $M_i(u)$ with $\hat{M}_i(u)$ for each i in equation (4.2). Important members of this class are the p score residuals for each participant, where p denotes the dimension of β . These residuals are defined by

$$L_{ij} \equiv \int_0^\infty H_{ij}(t) d\hat{M}_i(t), \quad i = 1, \dots, n, \quad j = 1, \dots, p,$$

where the left-continuous process $H_{ij}(t)$ is chosen such that $\sum_i L_{ij}$ reduces to the j th component of the partial likelihood score statistic. These p score residuals can be used to assess the influence of each participant on the parameter estimates $\hat{\beta}_j$ ($j = 1, \dots, p$). They are also related to a class of residuals, proposed by Schoenfeld (1982), that are useful for detecting departures from the proportional hazards assumption. Grambsch and Therneau (1994) explored the use of plots of scaled Schoenfeld residuals against event times to estimate the functional form of the hazard ratio over time.

Lin, Wei, and Ying (1993) studied the cumulative sums of martingale-based residuals over covariates or event times. The distributions of these stochastic processes under the assumed model can be approximated by zero-mean Gaussian processes. Each observed process can then be compared, both graphically and numerically, with a number of realizations from the approximate null distribution by computer simulation. These comparisons enable one to determine objectively whether a seemingly abnormal residual pattern reflects model misspecification or natural random variation.

10. Alternatives to the Cox Model

Despite the great popularity and versatility of the Cox regression model, there are reasons to explore alternative models. First, the proportional hazards assumption may not be satisfied in some applications. Second, alternative models characterize different aspects of the associations between covariates

and survival time. In this section, we describe briefly some alternative semiparametric models.

In contrast to the proportional hazards model, the additive hazards model specifies that covariates have additive rather than multiplicative effects on the hazard function, i.e.,

$$\lambda(t | \mathbf{Z}) = \lambda_0(t) + \beta' \mathbf{Z}(t). \quad (10.1)$$

This model was discussed by Cox and Oakes (1984, p. 147), Thomas (1986), and Breslow and Day (1987, p. 182). Using the counting process martingale approach, Lin and Ying (1994) obtained closed-form estimators for the regression parameters β and the cumulative baseline hazard function $\Lambda_0(t)$.

Semiparametric transformation models take the form

$$h(T) = \beta' \mathbf{Z} + \epsilon, \quad (10.2)$$

where ϵ is a random error with a given distribution function F and h is a completely unspecified function. If F is the extreme value distribution, then (10.2) is the proportional hazards model. If F is the standard logistic function, then (10.2) is the proportional odds model, under which the hazard ratio approaches unity as time increases. This class of models was studied by Clayton and Cuzick (1986) and Dabrowska and Doksum (1988), and the proportional odds model was studied by Pettitt (1982), Bennett (1983), and Murphy, Rossini, and van der Vaart (1997). A significant breakthrough was made by Cheng, Wei, and Ying (1995), who provided simple and relatively efficient estimators of β for all members of (10.2).

The semiparametric accelerated failure time model takes the same form as (10.2) but with h specified, usually as $h(T) = \log T$, and ϵ unspecified. Various methods of estimation for this model were proposed in the late 1970s and early 1980s. Koul, Susarla, and Van Ryzin (1981) suggested including in the least-squares estimator only the uncensored survival times but weighting them by the inverted probabilities of being uncensored. The resulting estimator is highly inefficient, especially in the presence of heavy censoring; however, the underlying idea of weighting uncensored observations by their inverted probabilities of being uncensored, to be referred to as the inverse probability of censoring weighting (IPCW) technique, turns out to be extremely useful in many other contexts. In fact, the Cheng et al. (1995) estimators were based on this idea. A more efficient modification of the least-squares estimator was provided by Buckley and James (1979), which replaces the conditional expectations for the censored survival times by their estimates based on the Kaplan–Meier estimator of the residual lifetime distribution and which involves an iterative estimation scheme analogous to the EM algorithm (Dempster, Laird, and Rubin, 1977). Prentice (1978), on the other hand, showed how to adapt the rank estimation method for noncensored data to the censored data setting. The asymptotic properties of the Buckley–James and rank estimators were established in the early 1990s by Tsiatis (1990), Ritov (1990), Wei, Ying, and Lin (1990), and Lai and Ying (1991a,b).

11. Sequential Analysis

When the efficacy and safety of two interventions are being compared in a clinical trial, interim monitoring often is performed in order to safeguard the interests of study participants and to achieve improved efficiency. To maintain false positive and false negative error rates when performing such

monitoring requires knowledge of the joint distribution of test statistics that are applied repeatedly over time. A major challenge in determining such distributions for censored data statistics is that there are two time axes: the survival time for each participant is measured from the time he/she enters the trial and sequential monitoring occurs over calendar time, which is measured from the start of the study.

Using the martingale theory and other probability techniques, Tsiatis (1982), Slud and Wei (1982), Harrington, Fleming, and Green (1982), Slud (1984), and Gu and Lai (1991) established the joint distributions for sequentially computed weighted log-rank statistics. With certain choices of the weight function, these sequential statistics (asymptotically) have normal independent increments over the calendar time, i.e., have the same joint distribution as the cumulative sums of independent normal random variables. Thus, group sequential monitoring guidelines based on normal responses, such as those proposed by Pocock (1977), O'Brien and Fleming (1979), and Lan and DeMets (1983), can be used. For the (unweighted) log-rank statistic, the increment is proportional to the number of observed events.

Sequential properties of censored-data statistics under the Cox model were studied by Sellke and Siegmund (1983), Tsiatis, Rosner, and Tritchler (1985), Gu and Ying (1995), and Biliias, Gu, and Ying (1997). Such statistics again have the normal independent increment structure. These results enable one to adjust for covariates in treatment comparisons and to construct repeated confidence intervals (Jennison and Turnbull, 1989) for the hazard ratio between treatments.

12. Multivariate Failure Time Data

Under the multiplicative intensity model described in Section 8, the risk of a recurrent event for a participant is unaffected by earlier events that occurred to the participant unless time-dependent covariates that capture such dependence are included explicitly in the model. In clinical trials applications, the dependence structures are complex and the forms of time-dependent covariates are unknown. Furthermore, the inclusion of such time-dependent covariates that are part of the response results in biased estimation of the overall treatment effect. Thus, it would be desirable to model the marginal distribution of the recurrent event times while leaving the dependence structures unspecified.

It is particularly appealing to consider the cumulative mean function $\mu(t) \equiv E\{N^*(t)\}$, where $N^*(t)$ is the number of events that the participant has actually experienced by time t (in the absence of censoring). This function was first considered by Nelson (1988) and further studied by Lawless and Nadeau (1995). A number of authors (e.g., Pepe and Cai, 1993; Lawless, Nadeau, and Cook, 1997; Lin et al., 2000) studied the following regression models for the cumulative mean function:

$$E\{N^*(t) \mid \mathbf{Z}\} = \mu_0(t)e^{\beta' \mathbf{Z}}, \quad (12.1)$$

where $\mu_0(t)$ is an arbitrary baseline mean function and β is a set of regression parameters. If $N^*(t)$ is a (nonhomogeneous) Poisson process, then (12.1) is equivalent to the intensity model determined by (8.1). Although in general N^* is not a Poisson process, the maximum partial likelihood estimator for β of (8.1) remains consistent and asymptotically normal under (12.1). The covariance matrix, however, can no longer

be estimated by the inversed information matrix. A sandwich variance estimator has to be used instead.

In the one-sample case, $\mu(t)$ can be consistently estimated by the Nelson estimator. Under model (12.1), the baseline mean function $\mu_0(t)$ can be consistently estimated by the Breslow estimator, and the covariate-specific mean function can be estimated in a similar fashion (see Lin et al., 2000). It is particularly informative to display the estimated mean functions for different treatment arms and for specific covariate patterns.

In some clinical trials, each participant can potentially experience more than one type of event. Examples include the developments of physical symptoms or diseases in several organ systems (e.g., stroke and cancer) or in several members of the same organ system (e.g., eyes or teeth). Models such as (8.1) and (12.1) are not applicable since the multiple events on the same participant are of different natures and in fact may not even be ordered.

It is convenient to formulate the marginal distributions of the multiple event times through the proportional hazards models while leaving the dependence structures completely unspecified. Let K denote the number of potential events per participant. The hazard function for the k th event of the i th participant is postulated to take the form

$$\lambda(t \mid \mathbf{Z}_{ki}) = \lambda_{k0}(t)e^{\beta' \mathbf{Z}_{ki}(t)}, \quad k = 1, \dots, K, \quad i = 1, \dots, n, \quad (12.2)$$

where \mathbf{Z}_{ki} is the covariate vector for the i th participant with respect to the k th event, λ_{k0} ($k = 1, \dots, K$) are arbitrary baseline hazard functions, and β is a set of regression parameters. In some applications (e.g., an ophthalmologic study involving the left and right eyes), it is natural to impose the restriction that $\lambda_{10} = \dots = \lambda_{K0}$, whereas in others (e.g., the setting of multiple diseases), it is necessary to allow the λ_{k0} 's to be different.

If the event times were independent, the partial likelihood could be easily constructed for β of model (12.2). The resulting estimator turns out to be consistent and asymptotically normal even if the event times are correlated; however, a sandwich variance estimator is again needed to account for the intraclass dependence. This approach was pioneered by Wei, Lin, and Weissfeld (1989) and further developed by Lee, Wei, and Amato (1992), Liang, Self, and Chang (1993), and Cai and Prentice (1995), among others.

The marginal approach discussed above treats the dependence of related event times as a nuisance. An alternative approach is to explicitly formulate the nature of dependence by the so-called frailty. The term frailty was first introduced by Vaupel, Manton, and Stallard (1979) to illustrate the consequences of a lifetime being generated from several sources of variation. The use of frailty in bivariate survival time data was considered by Clayton (1978). Frailty models were studied extensively in the 1980s by Clayton and Cuzick (1985), Hougaard (1987), and Oakes (1989), among others. The frailty-model analog of (12.2) specifies that the hazard function for the k th event of the i th participant, given the frailty ν_i , takes the form

$$\lambda_{ki}(t \mid \mathbf{Z}_{ki}; \nu_i) = \nu_i \lambda_{k0}(t)e^{\beta' \mathbf{Z}_{ki}(t)}, \quad (12.3)$$

where ν_i ($i = 1, \dots, n$) are independent random variables.

Conditional on ν_i , the event times on the i th participant are assumed to be independent.

The parameter vector β has a population-average interpretation under model (12.2) and a subject-specific interpretation under model (12.3). Models (12.2) and (12.3) cannot hold simultaneously unless ν is a positive-stable variable. It is very challenging, both theoretically and computationally, to deal with frailty models such as (12.3). Major progress was made in the 1990s. In the special case of gamma frailty models, maximum likelihood estimation via the EM algorithm was studied by Nielsen et al. (1992), Murphy (1994, 1995), Andersen et al. (1997), and Parner (1998), among others. Recently, Therneau and Grambsch (2000) suggested a penalized likelihood method for gamma and log-normal frailty models.

Nonparametric estimation for the multivariate survival function is a fundamental problem in the analysis of multivariate failure time data. Using the IPCW technique, Lin and Ying (1993) developed a simple estimator for the special case where there is a common censoring time for all event times of the same participant. Estimation in the general setting has been studied by Dabrowska (1988), Prentice and Cai (1992), and van der Laan (1996), among others.

The occurrence of one event (e.g., death) may preclude the development of another (e.g., relapse of cancer). In some applications, such as cause-specific mortality studies, the participant can only experience one of several potential events. This type of data is referred to as competing risks. The simplest solution to this problem is to censor the event time of interest at the time of the competing events and then apply the standard survival analysis methods such as the log-rank test and Cox regression. The results pertain to the so-called cause-specific hazard function, which is given by (2.1) with U representing the time to the competing events.

An important limitation of the cause-specific hazard function is that the associated $S^\#(t)$ is not a survival function unless the cause of interest is independent of other risks and the other risks could be eliminated without altering the distribution of the cause of interest. Thus, in general, the Kaplan–Meier estimator does not pertain to the survival function or disease incidence. Special methods have been developed to estimate disease incidence functions (e.g., Gray, 1988; Pepe and Mori, 1993; Fine and Gray, 1999).

13. Dependent Censoring and Related Problems

Virtually all the methods presented in the previous sections require the assumption of independent censoring. As discussed in Section 2, this assumption is generally true for censoring due to staggered entry but may not hold for censoring caused by voluntary patient withdrawal. As also discussed in Section 2, the survival distribution is not identifiable in the presence of dependent censoring. Lin, Robins, and Wei (1996) demonstrated that, for certain dependently censored data, it is possible to estimate the treatment difference in the survival distribution even if the survival distributions themselves are not estimable. If one is willing and able to model dependent censoring through time-dependent covariates, then the survival distribution is identifiable. Assuming that such time-dependent covariates are available, Robins and Rotnitzky (1992) showed how to use the IPCW technique mentioned in Section 10 to estimate the survival distribution.

Another type of dependent censoring arises in the contexts of quality-adjusted lifetime and medical cost. Because a participant who accumulates medical costs at higher rates tends to generate greater cumulative medical costs at both the survival time and censoring time than a participant with lower accumulation rates, the lifetime cost is positively correlated with its censoring variable, i.e., the cumulative cost at the censoring time (Lin, 2000). A similar phenomenon occurs for the quality-adjusted lifetime (Gelber, Gelman, and Goldhirsch, 1989). Thus, standard survival analysis methods, such as the Kaplan–Meier estimator and Cox regression, cannot be applied to censored medical costs or quality-adjusted lifetime data. The IPCW technique can be used to handle this type of dependent censoring (see Zhao and Tsiatis, 1997; Lin, 2000).

The IPCW technique turns out to be extremely useful for a very general type of censored/missing data satisfying coarsening at random (Heitjan and Rubin, 1991; Gill, van der Laan, and Robins, 1997). Under this assumption, the risk of censoring at time t depends on the full data (i.e., the data that would be observed in the absence of censoring) only through the history of observed time-dependent covariates up to t . For such data, the IPCW technique can be used to generate locally efficient estimators with double robustness in that the estimators are consistent if either the model for the full data or the model for the censoring mechanism is correctly specified and are asymptotically efficient if both models are correct. The local efficiency and double robustness were established, respectively, by Robins and Rotnitzky (1992) and by Scharfstein, Rotnitzky, and Robins (1999) in their rejoinder.

14. Additional Developments and Future Directions

Many areas of survival analysis have been reviewed in the previous sections. Further developments are anticipated in many of those areas. For example, it would be highly useful to develop efficient and reliable numerical algorithms for the semiparametric estimation of the accelerated failure time model. The Cheng et al. (1995) estimators for (10.2) require modeling the censoring distribution, and it would be worthwhile to explore estimation procedures that do not involve such modeling. In the area of multivariate failure time data, efficient estimators for model (12.2) have yet to be identified, and further theoretical and numerical advances are warranted for model (12.3). The theory of locally efficient estimators discussed in Section 13 offers tremendous power for solving a wide range of problems.

There are many other important areas of survival analysis that have not been covered in this article. In the remainder of this section, some are briefly mentioned, focusing on those areas in which future research activities are anticipated.

When the event of interest is asymptomatic, as is the case with cancer detection or HIV infection, the event time cannot be measured exactly but is rather known to lie in an interval determined by two successive examinations. Such data are said to be interval censored. The special case in which there is only a single examination time for each participant is called current status data. Parametric inferences for interval-censored data are straightforward, whereas non- and semiparametric inferences are very difficult. Significant progress has been made toward non- and semiparametric analysis of current status data. In particular, Groeneboom and Wellner

(1992) ascertained the limiting distribution of the nonparametric maximum likelihood estimator of the survival function. Huang (1996), Lin, Oakes, and Ying (1998), Rabinowitz, Tsiatis, and Aragon (1995), and Rossini and Tsiatis (1996) provided semiparametric methods under the proportional hazards, additive hazards, accelerated failure time, and proportional odds models, respectively. No such results are available for general interval-censored data, although some *ad hoc* methods (e.g., Finkelstein, 1986; Satten, 1996) have been suggested. Considerable future activities are expected in this area.

Clinical trial participants may not comply with the treatments they were randomly assigned to receive. In long-term follow-up studies, participants may change their treatments during the course of a trial according to their past responses. In fact, such dynamic treatment regimes are desirable from both the scientific and ethical points of view. Noncompliance and time-dependent treatments pose serious challenges. The use of time-dependent covariates to represent the treatments the participants actually received does not enable unbiased evaluation of treatments. Structural nested failure time models such as that of Robins and Tsiatis (1991) allow causal interpretation of treatment differences under certain conditions. Further research in this area is needed.

Failure time data is one type of data collected in long-term follow-up studies. Another type is so-called longitudinal data, which measures immediate responses, such as biological markers and quality of life scores, repeatedly over time. Joint modeling of longitudinal and failure time data has recently received tremendous attention from researchers in the fields of survival analysis and longitudinal data analysis. Significant progress is expected to occur in this area over the next few years.

The field of survival analysis has been dominated by the frequentist philosophy. The Bayesian approach has important potential in survival analysis, especially for handling complex models and data structures, such as dependent censoring, noncompliance, interval censoring, and multivariate failure times. Considerable progress has already been made in this area (see Sinha and Dey (1997) for an excellent review). One expects to see further developments and refinements of the Bayesian methods as well as their applications in clinical trials over the coming decades.

The applications of survival analysis methods to clinical trials have been greatly facilitated by the developments of software packages. Standard methods such as the Kaplan-Meier estimator, weighted log-rank tests, Cox regression, and parametric regression with (univariate) right-censored data are now available in virtually all software packages. The multiplicative intensity model and the sandwich variance estimators for models (12.1) and (12.2) have been implemented in major packages, such as SAS, S-Plus, and STATA. The penalized likelihood method for gamma and log-normal frailty models has recently been incorporated into S-Plus. However, most of the newer methods, such as those for the semiparametric analysis of models (6.1), (10.1), and (10.2) and those mentioned in this section, are not available in software packages. One is certainly to see further expansion of software.

Significant progress has also been made in other areas, such as adaptive tests, multistate models, time-dependent regression coefficients, nonparametric regression, graphical and

computer-intensive methods, surrogate markers, and missing/mismeasured covariates. Additional activities are anticipated in those areas.

The field of survival analysis was born and developed during the 20th century, partly in response to the needs in clinical trials applications. As evident from the above discussion, the field is still very much alive today and is expected to expand in the 21st century. Although it is difficult to predict exactly how the field will evolve in the future, it is certain that the interplay of survival analysis and clinical trials will continue to further the advancements of both fields in the new millennium.

ACKNOWLEDGEMENTS

This research was supported by the NIH grants CA42326, AI291968, and GM47845. The authors are grateful to the past editors Raymond Carroll and Louise Ryan for the invitation to write this article and to Diane Ames and Jane Fox Fulcher for their assistance in preparing the manuscript.

RÉSUMÉ

L'analyse de la survie est née au vingtième siècle, et a connu une importante croissance dans la seconde moitié du siècle. Les développements dans ce domaine qui ont eu le plus profond impact sur les essais cliniques sont la méthode de Kaplan-Meier (1958) pour l'estimation de la fonction de survie, la statistique du log-rank (Mantel, 1956) pour comparer deux distributions de survie, et le modèle des hasards proportionnels (Cox, 1972) pour quantifier les effets de covariables sur le temps de survie. La théorie des martingales pour processus de comptage défrichée par Aalen (1975) offre un cadre unifié pour l'étude des propriétés des statistiques d'analyse de survie, aussi bien pour petits que pour grands échantillons. Des progrès significatifs ont été réalisés, et on peut espérer de nouveaux développements dans plusieurs domaines, comme le modèle à temps accéléré, les données multidimensionnelles de survie, les données censurées par intervalle, la censure dépendante, les protocoles de traitement dynamiques et l'inférence causale, la modélisation jointe de données longitudinales et de données de durée, et les méthodes bayésiennes.

REFERENCES

- Aalen, O. O. (1975). Statistical inference for a family of counting processes. Ph.D. dissertation, University of California, Berkeley.
- Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics* **6**, 701–726.
- Aitkin, M. and Clayton, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics* **29**, 156–163.
- Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**, 1100–1120.
- Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. London: Springer-Verlag.
- Andersen, P. K., Klein, J. P., Knudsen, K. M., and Palacios, R. T. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics* **53**, 1475–1484.

- Anderson, G. L. (1989). Mismodelling covariates in Cox regression. Ph.D. dissertation, University of Washington, Seattle.
- Anderson, G. L. and Fleming, T. R. (1995). Model misspecification in proportioned hazards regression. *Biometrika* **82**, 527–541.
- Bailey, K. R. (1979). The general ML approach to the Cox regression model. Ph.D. dissertation, University of Chicago.
- Bailey, K. R. (1983). The asymptotic joint distribution of regression and survival parameter estimates in the Cox regression model. *The Annals of Statistics* **11**, 39–58.
- Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika* **75**, 65–74.
- Bartholomew, D. J. (1957). A problem in life testing. *Journal of the American Statistical Association* **52**, 350–355.
- Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273–277.
- Berkson, J. and Gage, R. (1950). Calculation of survival rates for cancer. *Proceedings of the Staff Meetings of the Mayo Clinic* **25**, 270–286.
- Biliyas, Y., Gu, M., and Ying, Z. (1997). Towards a general asymptotic theory for the Cox model with staggered entry. *The Annals of Statistics* **25**, 662–682.
- Breslow, N. E. (1972). Contribution to the discussion on the paper by D. R. Cox cited below. *Journal of the Royal Statistical Society, Series B* **34**, 216–217.
- Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.
- Breslow, N. E. and Crowley, J. J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics* **2**, 437–453.
- Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research, Volume 2, The Design and Analysis of Cohort Studies*. Lyon, France: IARC.
- Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–436.
- Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**, 151–164.
- Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.
- Chiang, C. L. (1968). *Introduction to Stochastic Processes in Biostatistics*. New York: Wiley.
- Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151.
- Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, Series A* **148**, 82–117.
- Clayton, D. G. and Cuzick, J. (1986). The semiparametrics Pareto model for regression analysis of survival times. In *Proceedings of the International Statistical Institute*, 23.3-1–23.3-18. Amsterdam: International Statistical Institute.
- Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society, Series B* **21**, 411–421.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.
- Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.
- Cox, D. R. and Snell, E. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B* **30**, 248–275.
- Crowley, J. J. (1970). A comparison of several life table estimates. Master's thesis, University of Washington, Seattle.
- Cutler, S. and Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival. *Journal of Chronic Diseases* **8**, 699–712.
- Cuzick, J. (1985). Asymptotic properties of censored linear rank tests. *The Annals of Statistics* **13**, 133–141.
- Dabrowska, D. M. (1988). Kaplan–Meier estimate on the plane. *The Annals of Statistics* **16**, 1475–1489.
- Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics* **15**, 1–23.
- Dempster, A. P., Laird, N. M., and Rubin, D. (1977). Maximum likelihood estimation for incomplete data via the EM algorithms (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
- Efron, B. (1967). The two-sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **4**, 831–853.
- Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* **72**, 557–565.
- Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* **19**, 69–76.
- Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**, 496–509.
- Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845–854.
- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Fleming, T. R. and Harrington, D. P. (1981). A class of hypothesis tests for one and two samples of censored survival data. *Communications in Statistics* **10**, 763–794.
- Fleming, T. R. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.
- Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1980). Modified Kolmogorov–Smirnov test procedures with application to arbitrarily right censored data. *Biometrics* **36**, 607–626.
- Fleming, T. R., Harrington, D. P., and O'Sullivan, M. (1987). Supremum versions of the logrank and generalized Wilcoxon statistics. *Journal of the American Statistical Association* **82**, 312–320.
- Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regression and omitted covariates. *Biometrika* **71**, 431–444.

- Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* **52**, 203–223.
- Gelber, R. D., Gelman, R. S., and Goldhirsch, A. (1989). A quality of life oriented endpoint for comparing therapies. *Biometrics* **45**, 781–795.
- Gill, R. D. (1980). *Censoring and Stochastic Integrals*, Volume 24, *Mathematical Centre Tracts*. Amsterdam: Mathematisch Centrum.
- Gill, R. D. (1983). Large sample behavior of the product-limit estimator on the whole line. *The Annals of Statistics* **11**, 49–58.
- Gill, R. D., van der Laan, M. J., and Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures and counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, D. Y. Lin and T. R. Fleming (eds), 255–294. New York: Springer-Verlag.
- Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526.
- Grambsch, P. M., Therneau, T. M., and Fleming, T. R. (1998). Residuals for survival analysis. In *Encyclopedia of Biostatistics*, Volume 5, P. Armitage and T. Colton (eds), 3813–3819. New York: Wiley.
- Gray, R. J. (1988). A class of K -sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics* **16**, 1141–1154.
- Greenwood, M. (1926). The natural duration of cancer. In *Reports on Public Health and Medical Subjects*, Volume 33, 1–26. London: Her Majesty's Stationery Office.
- Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel: Birkhäuser.
- Gross, A. J. and Clark, V. A. (1975). *Survival Distributions: Reliability Applications in the Biomedical Sciences*. New York: Wiley.
- Gu, M. and Ying, Z. (1995). Group sequential methods for survival data using partial likelihood score processes with covariate adjustment. *Statistica Sinica* **5**, 793–804.
- Gu, M. G. and Lai, T. L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *The Annals of Statistics* **19**, 1403–1433.
- Hájek, J. and Sidák, Z. (1967). *Theory of Rank Tests*. New York: Academic.
- Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553–566.
- Harrington, D. P., Fleming, T. R., and Green, S. J. (1982). Procedures for serial testing in censored survival data. In *Survival Analysis*, J. Crowley and R. A. Johnson (eds), 269–286. Hayward, California: Institute of Mathematical Statistics.
- Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics* **19**, 2244–2253.
- Hougaard, P. (1987). Modelling multivariate survival. *Scandinavian Journal of Statistics* **14**, 291–304.
- Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics* **24**, 540–568.
- Jacobsen, M. (1984). Maximum likelihood estimation in the multiplicative intensity model. *International Statistics Review* **52**, 193–207.
- Jennison, C. and Turnbull, B. W. (1989). Interim analyses: The repeated confidence interval approach (with discussion). *Journal of the Royal Statistical Society, Series B* **51**, 305–361.
- Johansen, S. (1983). The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics* **5**, 165–174.
- Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60**, 267–278.
- Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.
- Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right censored data. *The Annals of Statistics* **9**, 1276–1288.
- Lagakos, S. W. and Schoenfeld, D. A. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* **40**, 1037–1048.
- Lai, T. L. and Ying, Z. (1991a). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics* **19**, 531–556.
- Lai, T. L. and Ying, Z. (1991b). Large sample theory of a modified Buckley–James estimator for regression analysis with censored data. *The Annals of Statistics* **19**, 1370–1402.
- Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association* **76**, 231–240.
- Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.
- Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: Wiley.
- Lawless, J. F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**, 158–168.
- Lawless, J. F., Nadeau, C., and Cook, R. J. (1997). Analysis of mean and rate functions for recurrent events. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, D. Y. Lin and T. R. Fleming (eds), 37–49. New York: Springer-Verlag.
- Lee, E. W., Wei, L. J., and Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J. P. Klein and P. K. Goel (eds), 237–247. Dordrecht: Kluwer Academic.
- Liang, K. Y., Self, S. G., and Chang, Y. C. (1993). Modelling marginal hazards in multivariate failure time data. *Journal of the Royal Statistical Society, Series B* **55**, 441–453.
- Lin, D. Y. (2000). Linear regression analysis of censored medical costs. *Biostatistics* **1**, 35–47.
- Lin, D. Y. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* **80**, 573–581.

- Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.
- Lin, D. Y., Wei, L. J., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.
- Lin, D. Y., Robins, J. M., and Wei, L. J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* **83**, 381–393.
- Lin, D. Y., Oakes, D., and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika* **85**, 289–298.
- Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B*, **62**, 711–730.
- Liu, P. Y. and Crowley, J. (1978). *Large sample theory of the MLE based on Cox's regression model for survival data*. Technical Report 1, Wisconsin Clinical Cancer Center, Biostatistics, University of Wisconsin–Madison.
- Mann, N. R., Schafer, R. E., and Singpurwalla, N. D. (1974). *Methods for Statistical Analysis of Reliability and Life Data*. New York: Wiley.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Report* **50**, 163–170.
- McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.
- Mehrotra, K. G., Michalek, J. E., and Mihalko, D. (1982). A relationship between two forms of linear rank procedures for censored data. *Biometrika* **69**, 674–676.
- Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics* **22**, 712–731.
- Murphy, S. A. (1995). Asymptotic theory for the frailty model. *The Annals of Statistics* **23**, 182–198.
- Murphy, S. A., Rossini, A. J., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* **92**, 968–976.
- Naes, T. (1982). The asymptotic distribution of the estimator for the regression parameter in Cox's regression model. *Scandinavian Journal of Statistics* **9**, 107–115.
- Nelson, W. B. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology* **1**, 27–52.
- Nelson, W. B. (1988). Graphical analysis of system repair data. *Journal of Quality Technology* **20**, 24–35.
- Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19**, 25–43.
- Oakes, D. (1977). The asymptotic information in censored survival data. *Biometrika* **64**, 441–448.
- Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**, 487–493.
- O'Brien, P. C. and Fleming, T. R. (1979). A multi-stage procedure for clinical trials. *Biometrics* **35**, 549–556.
- Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics* **26**, 183–214.
- Pepe, M. S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88**, 811–820.
- Pepe, M. S. and Fleming, T. R. (1991). Weighted Kaplan–Meier statistics: Large sample and optimality considerations. *Journal of the Royal Statistical Society, Series B* **53**, 341–352.
- Pepe, M. S. and Mori, M. (1993). Kaplan–Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine* **12**, 737–751.
- Peterson, A. V. (1976). Bounds for a joint distribution function with fixed sub-distribution functions: Applications to competing risks. *Proceedings of the National Academy of Science, USA* **73**, 11–13.
- Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A* **135**, 185–206.
- Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *Journal of the Royal Statistical Society, Series B* **44**, 234–243.
- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.
- Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**, 167–179.
- Prentice, R. L. and Cai, J. (1992). Covariance and survival function estimation using censored multivariate failure time data. *Biometrika* **79**, 495–512.
- Rabinowitz, D., Tsiatis, A., and Aragon, J. (1995). Regression with interval censored data. *Biometrika* **82**, 501–513.
- Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics* **18**, 303–328.
- Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues*, N. P. Jewell, K. Dietz, and V. T. Farewell (eds), 297–331. Boston: Birkhäuser.
- Robins, J. M. and Tsiatis, A. (1991). Correcting for noncompliance in randomized trials using rank preserving structural equation failure time model. *Communications in Statistics—Theory and Methods* **20**, 2609–2631.
- Rossini, A. and Tsiatis, A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association* **91**, 713–721.
- Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* **83**, 355–370.
- Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* **94**, 1096–1146.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239–241.
- Schumacher, M. (1984). Two-sample tests of Cramer–von Mises and Kolmogorov–Smirnov type for randomly censored data. *International Statistics Review* **52**, 263–281.

- Self, S. G. and Prentice, R. L. (1982). Commentary on Andersen and Gill's Cox regression model for counting processes: A large sample study. *The Annals of Statistics* **10**, 1121–1124.
- Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika* **70**, 315–326.
- Sen, P. K. (1981). The Cox regression model, invariance principles for some induced quantile processes, and some repeated significance tests. *The Annals of Statistics* **9**, 109–121.
- Sinha, D. and Dey, D. K. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* **92**, 1195–1212.
- Slud, E. V. (1984). Sequential linear rank tests for two-sample censored survival data. *The Annals of Statistics* **12**, 551–571.
- Slud, E. V. and Wei, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistics. *Journal of the American Statistical Association* **77**, 862–868.
- Struthers, C. A. (1984). Asymptotic properties of linear rank tests with censored data. Ph.D. dissertation, University of Waterloo.
- Tarone, R. E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156–160.
- Therneau, T. M. and Grambsch, P. M. (2000). *Modelling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.
- Therneau, T., Grambsch, P., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147–160.
- Thomas, D. C. (1986). Use of auxiliary information in fitting nonproportional hazards models. In *Modern Statistical Methods in Chronic Disease Epidemiology*, S. H. Moolgavkar and R. L. Prentice (eds), 197–210. New York: Wiley.
- Tsiatis, A. A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Science, USA* **72**, 20–22.
- Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *The Annals of Statistics* **9**, 93–108.
- Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* **77**, 855–861.
- Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics* **18**, 354–372.
- Tsiatis, A. A., Rosner, G. L., and Tritchler, D. L. (1985). Group sequential tests with censored survival data adjusting for covariates. *Biometrika* **72**, 365–373.
- van der Laan, M. J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *The Annals of Statistics* **24**, 596–627.
- Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.
- Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**, 1065–1073.
- Wei, L. J., Ying, Z., and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**, 845–851.
- Wellner, J. A. (1982). Asymptotic optimality of the product limit estimator. *The Annals of Statistics* **10**, 595–602.
- Whitehead, J. (1980). Fitting Cox's regression model to survival data using GLIM. *Applied Statistics* **29**, 268–275.
- Zhao, H. and Tsiatis, A. A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* **84**, 339–348.

Received May 2000. Revised June 2000.

Accepted July 2000.