Wiley StatsRef:
Statistics Reference Online

Survival Analysis

By *Per K. Andersen*[1] *and Michael Vaeth*[2]

Keywords: *censoring, competing risks, Cox regression model, hazard rate, Kaplan–Meier estimator, logrank test, survival function, time scales*

Abstract: Survival analysis is a comprehensive set of statistical methods for analysis of data with incomplete observation of the main outcome variable, which is a time to occurrence of some specific event. The terminology reflects that many of the methods have been developed in connection with analysis of medical follow-up data where patients have been followed from a given date, typically day of treatment or day of diagnosis and until death. Such data are often censored, that is, the survival time is only known to exceed a given value, but other forms of incomplete data may also be encountered. Survival analysis includes nonparametric, semiparametric, and fully parametric models for comparisons of two or several groups and regression models that allow for correction of the influence of concomitant variables. In particular, the Cox proportional hazards regression models has become widely used for analysis of survival data. This model can be extended to include covariables that are updated during follow-up and a dynamic description of the problem can therefore be developed. Classical survival analysis includes a single event that each individual experiences at most once, but the methodology has been extended to cover much more complicated scenarios including data with several types of events and data with recurrent events. This entry surveys the most important aspects of survival analysis and presents a medical example to illustrate the use of some of the methods.

1 Survival Data

Survival analysis is concerned with statistical methods for analysis of data representing life times, waiting times, or more generally times to the occurrence of some specified event, the time being measured from some well-defined time origin until the event in question. Such data, denoted as *survival data*, can arise in various scientific fields and application of survival analysis methodology is commonly seen in medical studies and also in studies in social sciences and in industry. In a *clinical trial*, the object of the study may be the comparison of survival times with different treatments in some chronic disease; a social scientist may be interested in studying duration of employment or unemployment; in an industrial life testing experiment, a number of items could be put on test simultaneously and observed until failure. Thus, survival data are

[1] University of Copenhagen, Copenhagen K, Denmark
[2] Aarhus University, Aarhus, Denmark

Update based on original article by Kragh Andersen, Wiley StatsRef: Statistics Reference Online, © 2014, John Wiley & Sons, Ltd

DOI: 10.1002/9781118445112.stat02177.pub2

basically nothing but realizations of nonnegative random variables, but the statistical inference is usually complicated by the presence of incomplete observations. The most common form of incomplete survival data is *right censoring*, which reflects that limitations in time and other restrictions on data collection prevent the experimenter from observing the event in question for every individual or item under study. Rather, for some individuals, only partial information will be available about their survival time, namely, that it exceeds some observed *censoring* time, whereas the actual survival time is not observed or cannot be observed. The survival analysis methodology aims at estimating aspects of the distribution of the complete data from observation of the incomplete data. The mechanisms generating the incompleteness of the data must satisfy certain conditions of "independent censoring" to ensure that such inference is feasible. The presentation here will mainly consider incomplete data formed by right censoring and an adequate model of the censoring mechanism will depend on aspects of the particular problem to be described and the design of the study. A well-established way of formulating conditions for independent censoring is the following Ref. [1, p. 194]:

For an individual alive and *uncensored* at time t, the conditional probability of failing in the interval $[t, t + dt)$ given all that has happened up to time t coincides with the conditional probability of failing in $[t, t + dt)$ given survival up to time t.

An implication of independent censoring is the exclusion of censoring mechanisms withdrawing individuals from risk when they appear to have particularly high or low risk of failure. [For alternative formulations of independent censoring, see, e.g., Ref. [2, p. 139] (*see* **Censored Data and Clinical Trials**)]. If a survival data model includes covariates, $\mathbf{x}$ (see the following discussion) then the condition should hold for any given value of $\mathbf{x}$.

In a clinical trial, patients are typically followed from some well-defined event, for example, diagnosis, treatment, or randomization, until the endpoint occurs or the patient is censored alive. The relevant timescale is time since entry and all patients are therefore followed from time 0. This is, however, not always the case with other types of study designs. Individuals with a chronic disease may be identified on a given date and then followed forward in time until death. The relevant timescale could here be time since diagnosis, but patients are not followed from the date of diagnosis: To be included in the sample, a patient must survive until the date that the sample is identified. This type of incomplete observation is denoted left truncation or *delayed entry* because entry occurs after time 0 and patients are excluded if they experience the event before entry. Statistical methods developed for analysis of right censored data may often be modified to allow for both left truncation and right censoring provided that the truncation mechanism satisfies a condition of "independent truncation" similar to that of "independent censoring" (*see* **Delayed Entry**).

The object of a survival analysis is to draw inferences about the distribution of the survival times T. This distribution can be characterized by the survival distribution function (sdf),

$$S(t) = \Pr[T > t], t \geq 0$$

or equivalently (provided that S is differentiable) by the *hazard rate* function (hrf)

$$\lambda(t) = -d(\log\ S(t))/dt = f(t)/S(t), t \geq 0 \tag{1}$$

where $f(t)$ is the probability density function (pdf) or the *cumulative hazard* rate function (chrf)

$$\Lambda(t) = \int_0^t \lambda(u)du, t \geq 0 \tag{2}$$

Note that the hrf defined in Equation (1) (when $dt > 0$ is small) has the following attractive "dynamic" interpretation:

$$\lambda(t)dt \approx \Pr[T \leq t + dt \mid T > t]$$

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat02177.pub2

Statistical models for continuous survival data can be specified via any of these quantities but it has become customary to formulate these models in terms of the hrf. In particular, the influence of *covariates* on the survival time is often specified via the hrf.

In survival analysis, focus is usually not on the mean survival time E[T]. This has to do with the fact that the mean obviously depends heavily on the right-hand tail of the distribution and that right censoring often leaves the researcher with little information on that tail.

Throughout it will be assumed that the censoring and the truncation mechanisms are independent and that the available data are of the form of times of observation $t_1, \ldots, t_n$ and indicators $d_1, \ldots, d_n$. The latter are used to distinguish between uncensored and censored observations, that is, $d_i = 1$ if t_i is an actual survival time and $d_i = 0$ if t_i is a censored observation. Delayed entry is included in this set-up by adding an entry time v_i satisfying $0 \leq v_i < t_i$ for each individual. For methods where $S(t)$ is discontinuous, corresponding to survival distributions with discrete components, see Ref. [1, p.46] (*see* **Discrete Survival-Time Models**). Methods for dealing with grouped survival data do exist ; the classical actuarial *life table* is the main example (*see* **Life Table: Overview**).

2 Nonparametric Survival Models

Presence of censored observations implies that classical nonparametric methods based on *ranks* are not directly applicable. Moreover, standard graphical procedures such as an *empirical cdf* or a *histogram* cannot be used with censored data. To present the required modifications, it is convenient to consider the following quantities defined from the basic observations $(t_i, d_i, v_i), i = 1, \ldots, n$:

$$N(t) = \#\{i = 1, \ldots, n : v_i < t_i \leq t, d_i = 1\}$$

and

$$Y(t) = \#R(t),$$

where

$$R(t) = \{i = 1, \ldots, n : t_i \geq t > v_i\}$$

is the *risk set* at t. Thus, $N(t)$ is the *counting process* counting the number of failures observed before or at t and $Y(t)$ is the number of individuals observed to be *at risk* at $t-$ (that is, just before t). The nonparametric methods to be discussed in this section have all been reviewed in Ref. [2, Chapter 4 and 5] (*see* **Counting Process Methods in Survival Analysis**).

2.1 Estimation

The sdf can be estimated by the *Kaplan–Meier* or product limit estimate[3]

$$\hat{S}(t) = \prod_{v_i < t_i \leq t} \left[1 - \frac{\Delta N(t_i)}{Y(t_i)}\right],$$

where $\Delta N(t) = N(t) - N(t-)$ is the number of failures at t. Similarly, the chrf defined in Equation (2) can be estimated by the *Nelson–Aalen estimate*[4]

$$\hat{\Lambda}(t) = \sum_{v_i < t_i \leq t} \frac{\Delta N(t_i)}{Y(t_i)}.$$

DOI: 10.1002/9781118445112.stat02177.pub2

Conditions (including independent censoring) can be found under which $\hat{S}(t)$ and $\hat{\Lambda}(t)$ behave asymptotically ($n \to \infty$) as normal processes, and approximate standard errors can be calculated as

$$\hat{\sigma}(\hat{\Lambda}(t)) = \left[\sum_{v_i < t_i \le t} \frac{\Delta N(t_i)}{Y(t_i)(Y(t_i) - \Delta N(t_i) + 1)} \right]^{1/2},$$

$$\hat{\sigma}(\hat{S}(t)) = \hat{S}(t)\hat{\sigma}(\hat{\Lambda}(t))$$

The estimate $\hat{\sigma}(\hat{S}(t))$ is known as *Greenwood's formula*. Estimates $\hat{\lambda}(t)$ of the hrf can be obtained, for example, by smoothing $\hat{\Lambda}(t)$ using some *kernel function*.

2.2 Hypothesis Tests

Nonparametric comparison of the survival distributions in $k = 2$ groups of individuals can be based on test statistics of the form

$$Z = \sum_{i=1}^{n} \left[K(t_i) \left\{ \frac{\Delta N_2(t_i)}{Y_2(t_i)} - \frac{\Delta N_1(t_i)}{Y_1(t_i)} \right\} \right]$$

where $K(t)$ is a stochastic "weight" process. Under suitable conditions $ZV^{-1/2}$, where

$$V = \sum_{i=1}^{n} \left\{ K^2(t_i) \frac{\Delta N_1(t_i) + \Delta N_2(t_i)}{Y_1(t_i) Y_2(t_i)} \right\}$$

has an asymptotic standard normal distribution (as $n \to \infty$). Different choices of K lead to test statistics discussed in the survival data literature. For example, the choice $K = Y_1 Y_2 / Y$ yields the *logrank test* where Z reduces to the difference between the observed $O_j = N_j(\infty)$ and the "expected" number of failures in group j(=1 or 2):

$$E_j = \sum_{i=1}^{n} \left[\frac{Y_j(t_i)}{Y(t_i)} \Delta N(t_i) \right]$$

Similar tests for comparison of the survival distributions in $k \ge 2$ groups of individuals and for comparing an observed survival distribution with an sdf $\exp(-\Lambda^*(t))$ that is known (for example, from the life tables for some reference population) are discussed in Ref. [2, Chapter 5] (*see* **Linear Rank Tests in Survival Analysis**).

3 Parametric Survival Models

Survival data are often skewed so the normal distribution plays a less prominent role in the analysis of such data. Popular parametric survival models include the *exponential distribution* and the *Weibull distribution*. The hrf has a simple form for these distributions and parameter estimation is simplified because a closed-form expression for the survival function is available. The basic properties of these distributions are reviewed below, the emphasis being on the aspects of the distributions that are important in survival analysis. Moreover, survival distributions with piecewise-constant hazard rate and the log-normal distribution will be briefly discussed. For further reading, see Ref. 5 (*see* **Parametric Models in Survival Analysis**).

3.1 Some Special Parametric Models

The simplest lifetime distribution, the exponential distribution, is characterized by a constant hrf $\lambda(t) = \lambda$ for all $t \ge 0$, and the sdf has the form $S(t) = \exp(-\lambda t)$. Although the assumption of a constant hazard

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat02177.pub2

rate is restrictive, the exponential distribution was the first survival model to become widely used, partly due to its computationally attractive features. The appropriateness of the exponential distribution for a given set of survival data may be checked by plotting $\hat{\Lambda}(t)$ or equivalently $-\log \hat{S}(t)$ versus t. Such a plot should approximate a straight line through the origin. With censored data from an exponential distribution, maximum likelihood estimation is particularly simple and the maximum likelihood estimate becomes the rate $\hat{\lambda} = D/\sum t_i$, where $D = N(\infty) = \sum_i d_i$ is the total number of events.

Improved flexibility without sacrificing the simplicity of the computations can be obtained by assuming that the hazard rate is piecewise constant on a number of prespecified time intervals, for example, 1 or 5-year intervals. With censored data from this distribution, the maximum likelihood estimate of the hazard rate λ_j on the jth interval becomes $\hat{\lambda}_j = D_j/T_j$, where D_j is the number of events in the jth interval and T_j the sum of the time periods each individual spends in the jth interval. From a computational point of view, a survival model with piecewise constant hazard is closely related to Poisson distributions and this connection may utilized when the model is generalized to include covariates, see the Section 4.2.

The Weibull distribution is probably the most widely used parametric survival model in applications. The Weibull model provides a fairly flexible class of distributions and includes the exponential distribution as a special case. The hrf and the sdf have the form

$$\lambda(t) = \lambda\rho(\lambda t)^{\rho-1}, \quad t \geq 0,$$

and

$$S(t) = \exp(-(\lambda t)^{\rho}), \quad t \geq 0 \tag{3}$$

where $\lambda > 0$ is an inverse scale parameter and $\rho > 0$ is a shape parameter. The hrf is monotonically increasing if $\rho > 1$, monotone decreasing if $0 < \rho < 1$, and constant if $\rho = 1$. The Weibull distribution appears as one of the asymptotic distributions of the smallest extreme and this fact motivates its use in certain applications. From Equation (3), it follows that

$$\log \Lambda(t) = \rho \log \lambda + \rho \log t$$

and the appropriateness of the Weibull model can therefore be checked by plotting $\log \hat{\Lambda}(t)$ versus $\log t$.

The lognormal distribution has also been used in survival analysis, although no closed-form expressions are available for $S(t)$ and $\lambda(t)$. The distribution is most easily specified through $\log T$ having a normal distribution with mean μ and variance σ^2, say, when T is a lognormal variate. The lognormal hrf has the value 0 at $t = 0$, increases to a maximum, and then decreases with a limiting value of 0 as t tends to infinity. This behavior may be unattractive in certain applications.

Note that a parametric specification of a survival distribution also implicitly specifies the mean survival time $E[T]$. However, focus on the mean will often involve an extrapolation beyond the range of the observed survival times.

3.2 Inference

For the parametric survival models, including the Weibull model and the lognormal model, maximum likelihood estimation and large-sample likelihood methods are the inference procedures generally used, as the presence of incomplete data makes exact distributional results complicated in most situations.

For the class of independent censoring schemes, the likelihood function becomes (apart from a constant of proportionality)

$$L(\theta) = \prod_{i=1}^{n} \lambda(t_i;\theta)^{d_i} S(t_i;\theta) \tag{4}$$

where θ is the vector of unknown parameters to be estimated. When delayed entry is also present in the data, the ith factor in Equation (4) is divided by $S(v_i;\theta)$. From Equation (4), it is apparent that survival models admitting a closed-form expression for the sdf are computationally attractive. In most cases, the

DOI: 10.1002/9781118445112.stat02177.pub2

standard errors of the parameter estimates have to be estimated from the observed information matrix since calculation of the expected information requires detailed knowledge of the distribution of the censoring times.

The theoretical justification of the asymptotic normality of the maximum likelihood estimate and the limiting χ^2 distribution of the likelihood ratio statistic with censored data has been established in many special cases. A unified approach to the asymptotic theory of maximum likelihood estimation for parametric survival models with independent censoring and truncation has been given in Ref. [2, Ch. 6].

4 Regression Models

Application of survival analysis usually involves evaluation of the dependence of the survival time on one or several concomitant variables (covariates) measured on each individual. Regression modeling is the standard approach to identify and evaluate such prognostic variables. Several regression models, which allow for right censoring, have been developed, the most important being *Cox's proportional hazards regression model*. In most of these regression models (including Cox's regression model), the influence of a prognostic variable on the survival time is modeled by letting the hrf depend on the prognostic variables.

4.1 The Cox Proportional Hazards Regression Model

Cox's regression model[6] is a semiparametric model in which the dependence on time is described by an unspecified reference hazard function and the influence of a prognostic variable is modeled by modifying this hazard function by a multiplier independent of time. The hrf for a subject with covariates $\mathbf{x} = (x_1, \ldots, x_p)$ can be written as

$$\lambda(t, \mathrm{x}) = \lambda_0(t) \exp(\beta' \mathrm{x}) \tag{5}$$

where β is a p-dimensional vector of unknown regression coefficients reflecting the effects of $\mathbf{x}$ on survival and $\lambda_0(t)$, the baseline hrf, is an unspecified function of time (*see* **Cox Regression Model**). The basic features of the model (Equation 5) are the assumption of *proportional hazards*, that is, $\lambda(t, \mathbf{x}_1)/\lambda(t, \mathbf{x}_2)$ does not depend on t and the assumption of log linearity in $\mathbf{x}$ of the hrf. The parameters of the model are most easily interpreted as log hazard ratios. In particular, for a prognostic factor x_1 with two categories exposed ($x_1 = 1$) and unexposed ($x_1 = 0$), $\exp(\beta_1)$ becomes the hazard ratio between an exposed individual and an unexposed individual, who is otherwise identical to the exposed individual. The statistical analysis of Cox's regression model is based on the partial likelihood function (*see* **Partial Likelihood**), which in the case of no ties between the survival times is given by

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' \mathbf{x}_i)}{\Sigma_{j \in R(t_i)} \exp(\beta' \mathbf{x}_j)} \right)^{d_i} \tag{6}$$

Here, $\mathbf{x}_i$ is the covariate vector of the individual with observation time t_i and $R(t_i)$ is the risk set at t_i. Cox's partial likelihood function (6) may be obtained from the likelihood function (4) by profiling out the baseline hazard function $\lambda_0(t)$. Estimates of the parameters β are obtained by maximizing $L(\beta)$ and the usual type of large-sample likelihood methods also apply to partial likelihoods when censoring is independent and certain regularity assumptions are satisfied; see, for example, Ref. [2, Ch. 7]. The chrf $\Lambda_0(t)$ corresponding to $\lambda_0(t)$ can, in the case of no tied survival times, be estimated by

$$\hat{\Lambda}_0(t) = \sum_{t_i \le t} \frac{d_i}{\sum\limits_{j \in R(t_i)} \exp(\hat{\beta}' \mathbf{x}_j)}$$

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat02177.pub2

the so-called Breslow estimator. When ties are present in the data, modifications of both Cox's partial likelihood and the Breslow estimator are available. In applications, the parameter estimates, as exemplified below, are usually presented as hazard ratios with 95% confidence intervals. These are obtained as the exponential function of $\hat{\beta}_j$ and of the 95% confidence limits for β_j. Regression diagnostics for Cox's regression model also include methods for checking the validity of the proportional hazards assumption. A general approach to model checking is based on cumulative martingale residuals, see, for example, Ref. 7. If a particular prognostic variable influences the survival time in a way that can not be described by proportional hazards, one may consider a stratified Cox's regression model. This model includes separate, not necessarily proportional, baseline hazard functions for each category of a prognostic variable, whereas the dependence on the remaining prognostic variables is described by proportional hazards (*see* **Residuals for Survival Analysis**; **Goodness of Fit in Survival Analysis**).

Cox's regression model can be generalized to allow for *time-dependent covariates*. This is an important extension that increases the flexibility of the model considerably. Three broad classes of time-dependent covariates can be identified. As an alternative to a Cox regression model stratified on a covariate, x, one may add a known function of time, for example, $x \cdot \log(t)$, as a covariate in order to describe an influence of x acting on survival in a nonproportional way. Such a model including an interaction between a covariate and a function of time, t, also provides a means for testing for proportional hazards. Another type of time-dependent covariate is used to model the influence of prognostic variables for which information is updated during follow-up. Rather than using the value available at entry, one may want to include the most recent value in the model. Finally, one may want the model to reflect the influence of events that may be observed during follow-up. This can be done by introducing a time-dependent binary covariate that takes the value 0 until the event occurs and the value 1 thereafter. Comprehensive reviews of the model are given by Kalbfleisch and Prentice Ref. [1, Chapters 4–6] and Andersen *et al.* Ref. [2, Chapter 7] (*see* **Time-Dependent Covariate**).

It should be noted that if random time-dependent covariates (e.g., of the second and third type just described) are included in the model for the hrf then the relation $S(t) = \exp(-\Lambda(t))$ implied by Equation (5) no longer holds. This is because the cumulative survival probability will also depend on the random future development of the time-dependent covariate. Consequently, effects of time-dependent covariates should be interpreted with caution because the effect of a covariate on the hrf may now differ from its effect on the sdf. In such a situation, joint models for the time development of the covariate and the hrf may be studied, see for example, Ref. 8. An alternative and much simpler approach is to use the so-called landmarking, Ref. 9. Here, a number of "landmark" time points (say, $\tau_1, \ldots, \tau_L$) are chosen, and at the jth landmark, the conditional distribution of the future life time given survival till τ_j is modeled considering the current values of covariates but without accounting for future changes in covariate values. Such models are useful for prediction purposes as now the relationship (Equation 1) between the hrf and the sdf still holds from the landmark time and onward.

4.2 Parametric Hazard Regression Models

Parametric proportional hazards models are obtained by replacing the arbitrary function $\lambda_0(t)$ by a function belonging to some parameterized family and then statistical inference is usually based on a likelihood function analogous to Equation (4). The simplest examples arise when $\lambda_0(t)$ is replaced by a constant (exponential regression) or by a power function of time (Weibull regression). Using a piecewise constant hrf, the likelihood function becomes proportional to a likelihood function obtained by formally treating the number of events in each interval as independent Poisson variates with mean equal to the hazard rate multiplied by the time at risk. This model can therefore be analyzed with software for *generalized linear models* and the model is usually called the *Poisson regression* model. If all covariates are categorical, the data can be aggregated into a multiway table of counts and person-years at risk. This may be a useful approach for analyzing data from large *cohort studies* and this Poisson model therefore has important applications when

DOI: 10.1002/9781118445112.stat02177.pub2

analyzing observational studies in epidemiology, for example, based on large population registers. A useful parametric alternative to the Cox model is obtained by modeling the baseline hazard as a flexible spline function and interactions between covariates and time, that is, nonproportional hazards, are also modeled using spline functions of time, see Ref. 10 for a comprehensive discussion of this regression model.

4.3 Additive Hazard Models

As an alternative to the multiplicative structure in Cox's regression model, additive hazard rate models have been developed, the most flexible one being *Aalen's additive hazard rate model*, for example, Ref. [2, Chapter 7, Ref. 11, Chapter 4 and Ref. 7, Chapter 5] (*see* **Aalen's Additive Regression Model**). In this model, the hrf for a subject with covariates $\mathbf{x} = (x_1, \ldots, x_p)$ (which may be time dependent) is given by

$$\lambda(t, \mathrm{x}) = \beta_0(t) + \beta(t)'\mathrm{x}.$$

Here, $\beta_0(t)$ is an unspecified baseline hrf and $\beta_j(t), j = 1, \ldots, p$, are unspecified regression functions quantifying the influence of the covariates. Thus, for a binary covariate, x_j, taking values 0 or 1, the value at time t of the corresponding regression function, $\beta_j(t)$, describes the hazard difference at t between subjects with $x_j = 1$ and $x_j = 0$ and all other covariates identical. Both the cumulative baseline hazard and the cumulative regression functions, $B_j(t) = \int_0^t \beta_j(u)du, j = 0, 1, \ldots, p$ can be estimated nonparametrically by "generalized *Nelson–Aalen estimators*" whose increments at a failure time t_i are obtained by solving a certain least squares equation at that time. Asymptotic inference including standard errors and test statistics for linear hypotheses of the form $c'\beta(t) = 0$ are also available Ref. [11, Chapter 4]. A simple special case is when all regression functions are constant:

$$\lambda(t, \mathrm{x}) = \beta_0(t) + \beta'\mathrm{x}$$

but where the baseline hazard is still unspecified. In addition, semiparametric versions where some regression functions are constant and others are left unspecified have been studied Ref. [7, Chapter 5].

4.4 The Accelerated Failure Time Model

The *accelerated failure time regression model* is conveniently introduced via the logarithm of the survival time T. The model specifies a linear relationship

$$\log T = \gamma'\mathrm{x} + \sigma W \tag{7}$$

where σ is a scale parameter and W a mean zero random variable giving the error. Thus, $\gamma'\mathrm{x}$ is the mean value of $\log(T)$. Various choices of the error distribution lead to regression versions of the parametric survival models already discussed. Specifically, if W has an *extreme value distribution* (*Gumbel distribution*), a Weibull regression model is obtained, the exponential regression being a special case corresponding to $\sigma = 1$.

A lognormal regression model is obtained if W is a standard normal variate. Parametric statistical inference for the model (7) based on a likelihood function analogous to Equation (4) has been described by Kalbfleisch and Prentice Ref. [1, Chapter 3] and Lawless Ref. [5, Chapter 5]. Nonparametric analysis of the model (Equation 7) has also been developed Ref. [2, Ch. 7] and Ref. [7, Chapter 8]. However, inference for such a model is complicated by right censoring that often prevents the right-hand tail of the survival time distribution to be estimated.

The accelerated failure time regression model can alternatively be formulated via the HRF, namely,

$$\lambda(t, \mathbf{x}) = \lambda_u[t \exp(-\boldsymbol{\gamma}'\mathbf{x})] \exp(-\boldsymbol{\gamma}'\mathbf{x})$$

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat02177.pub2

where $\lambda_u(\cdot)$ denotes the HRF of the random variable $U = \exp(\sigma W)$. The only such models that are also proportional hazards models are the exponential and the Weibull regression models.

4.5 Several Time scales

Some studies may involve several time variables, for example, age at diagnosis, current age, time since first diagnosis, time since last episode, and identification of the most important timescale may be part of the purpose of the analysis. Statistical methods in survival analysis (most prominently, the Cox regression model) often require a primary timescale and other timescales may then be introduced as time-dependent covariates. However, with Poisson regression, several time variables can be introduced and studied simultaneously without selection of a primary timescale.

5 Some Extensions

5.1 Interval Censoring

Right censoring and delayed entry are the most common forms of incomplete observation of survival data, but depending on the study design, other types of incomplete observation may also be encountered. If patients in a clinical trial are seen at scheduled visits to the clinic at regular intervals, the time of occurrence of some health-related event may not be identified precisely but only referred to a period between visits. This type of incomplete observation is denoted **interval censoring** (*see* **Interval Censoring: Overview**). The survival time is here only known to lie in some interval. *Grouped survival data* are a special case of interval censored data for which the possible intervals are the same for all individuals (*see* **Grouped Survival Times**). Current status data may also be considered as a special case of interval censoring; here the patient is examined only once and it is determined if the event has occurred or not at that particular point in time. In general, interval censoring requires special methodology; however, parametric inference based on likelihood functions is easily developed for interval censored data. If the survival time of an individual is known to lie in the interval from L_i to R_i, this person contributes with a factor of $S(L_i) - S(R_i)$ to the likelihood function. Nonparametric inference for interval censored data has also been developed, for example Ref. 12.

5.2 Multistate Models, Competing Risks

The models mentioned so far can be thought of as describing transitions from one state "alive" to another state "death," the transition rate being the HRF $\lambda(t)$. Many of the methods described can be extended to situations where more than one type of transition can occur, the so-called multi-state models. The simplest such extension is the *competing risks* model where the state "death" is split into, say, a states: "death from cause 1," … , "death from cause a" (*see* **Event History Analysis**). In this model, there are a transition intensities, the so-called cause-specific HRFs, $\lambda_j(t)$, $j = 1, \ldots, a$ corresponding to the different causes of death, each of which may be analyzed using, for example, hazard regression models [Ref. 2, Chapter 8]. In the competing risks model, there is still one SDF,

$$S(t) = P(T > t) = \exp\left(-\sum_{j=1}^{a}\int_0^t \lambda_j(u)du\right)$$

DOI: 10.1002/9781118445112.stat02177.pub2

giving the probability of surviving beyond time t. The CDF, $F(t) = 1 - S(t)$, however, is split into a sum of a "cumulative Incidences," $F(t) = \sum_{j=1}^{a} F_j(t)$, where

$$F_j(t) = P(T \leq t, D = j) = \int_0^t S(u-)\lambda_j(u)du$$

Here, D is the cause of death indicator. The cumulative incidence for cause j, $j = 1, \ldots, a$, is the transition probability from the initial state "alive" to the final state "death from cause j," that is, it gives the probability of failure from cause j before time t. Note that $F_j(t)$ (via $S(u)$) depends on the cause-specific hazards for *all* causes and it may be estimated nonparametrically by plugging-in Nelson–Aalen estimators for the cumulative cause-specific HRFs and the Kaplan–Meier estimator for the SDF Ref. [2, Chapter 4]. Similarly, regression models for the cumulative incidence may be obtained by plugging-in regression models for the cause-specific hazards. However, with this approach, the final model does not include parameters that directly describe the association between a covariate and the cumulative incidence. This may complicate the interpretation of the results, and alternative, direct regression models for cumulative incidence have therefore been developed. One such model is the Fine-Gray model while alternative models are based on pseudo-observations or the so-called direct binomial regression. These approaches are reviewed in Ref. 13.

In the illness-death model and other multistate models, plug-in models for the transition intensities are also a possibility, but for some such models, the transition probabilities are complicated functions of the transition intensities and this approach therefore becomes intractable Ref. [2, Chapter 4]. Regression approaches based on pseudo-observations or direct binomial models have also been developed for more general multistate models, see, for example, Ref. 13.

5.3 Multivariate Survival Data, Recurrent Events

In the survival data models described in previous sections, an implicit assumption has been *independence* between the n observations. However, *clustering* may appear both in the form of "serial" and "parallel" survival data. Serial data, or *recurrent events*, corresponds to the situation where, for the same subject, the event of interest could occur repeatedly, for example, episodes of a nonchronic disease (*see* **Repeated Events**). Examples of parallel data could be survival times observed in families, for example, twins, or in medical centers (*see* **Multivariate Survival Analysis**).

In both situations, independence within clusters is questionable while independence between clusters may still be a reasonable assumption and the survival data inference needs to address this potential intracluster dependence. As for other types of outcome variables, for example, quantitative or binary data, there are basically two types of approach for addressing the dependence. One, *random effects models*, explicitly describes the dependence by letting the outcome for subjects from the same cluster share common random effects, whereas the other, *marginal models*, treats the intracluster association as a nuisance parameter and aims at estimating model parameters (and standard errors) in a way which is robust to the lack of independence. Both types of approach have been studied for both serial and parallel survival data.

In random effects models for survival data, the HRFs for subjects, $j = 1, \ldots, n_i$ from the same cluster, i, are typically related by sharing a common, unobservable random cluster-specific "frailty" Z_i. Such models, known as *frailty models*, specify the HRF for individual j in cluster i by

$$\lambda_{ij}(t) = z_i \lambda_j(t) \tag{8}$$

when $Z_i = z_i$ and $\lambda_j(t)$ are the HRF for an individual with unit frailty Ref. [14, Chapter 7–10]. Models of the form (8) are analogous to variance components models for quantitative outcomes and inference typically assumes that $Z_1, \ldots, Z_m$, where m is the number of clusters, are independent and identically distributed with a known type of distribution. For mathematical convenience, the *gamma distribution* has played a prominent role for *frailty models* (*see* **Frailty**).

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat02177.pub2

In marginal models for clustered, parallel survival data, on the other hand, a model for the HRF of the marginal distribution for the survival time of subject j in cluster i is typically specified using one of the hazard-based models described earlier, such as the Cox regression model. Parameters are then estimated using *generalized estimating equations*, (GEEs), and standard errors, robust to lack of independence within clusters, are obtained using "sandwich-type" formulas Ref. [1, Chapter 10]. Marginal models for serial data Ref. [15, Chapter 3] are typically specified for the mean number of events, $E(N_i(t))$, for subject i, where $N_i(t)$ is the process counting the number of recurrent events for i before time t.

6 An Example

In the period 1964–1977, 205 patients with malignant melanoma had radical surgery performed at the Department of Plastic Surgery, University Hospital of Odense, Denmark Ref. [2, Chapter 1]. At the end of the follow-up period (1 January 1978), 57 of the patients had died from the disease and 14 patients had died from other causes. Figure 1 shows a stacked plot of the cumulative incidences for these two causes of death. The plot shows that after 5 years the estimated risk of dying from malignant melanoma is 22.4%, the risk of dying from other causes is estimated to 4.4%, and the chance of being alive is estimated to 73.2%. Among the variables registered in the study were the sex of the patient and the age at surgery, tumor thickness, and ulceration (absent or present).

The first analysis considers mortality from all causes with survival time measured from the date of operation. Figures 2 and 3 show the Kaplan–Meier estimates of the survival functions in groups defined by the sex of the patients and ulceration. Table 1 presents the results of two logrank analyses comparing the survival distributions in the groups. It is seen that, when considered separately, both sex and ulceration have a significant influence on survival. Cox regression analyses of each covariate separately confirm these findings. The estimated hazard ratios with 95% confidence intervals for sex (male relative to female) and ulceration (present relative to absent) were 1.94 (1.15; 3.26) and 4.36 (2.44; 7.77), respectively. Moreover, the estimated hazard ratios from single factor Cox regression analyses of age at surgery (per 10 years) and tumor thickness included as ln(thickness in *mm*) were 1.21 (1.02; 1.44) and 2.02 (1.55;2.64)

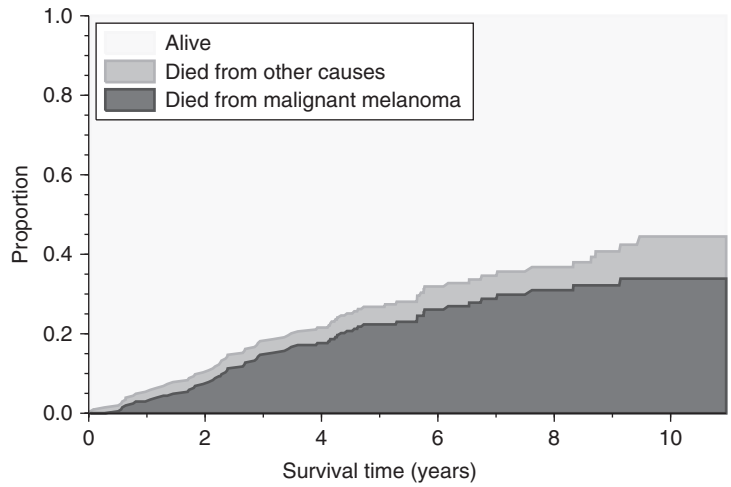

Figure 1. Stacked plot of the cumulative incidences for death from malignant melanoma and death from other causes.

DOI: 10.1002/9781118445112.stat02177.pub2

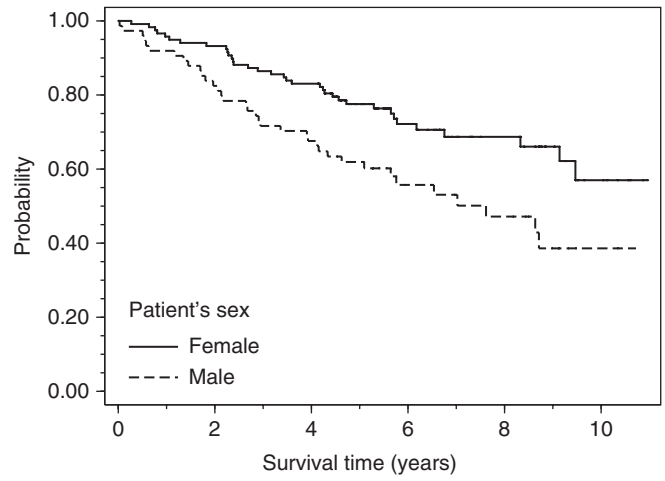

Figure 2. Kaplan–Meier estimates for male and female patients with malignant melanoma.

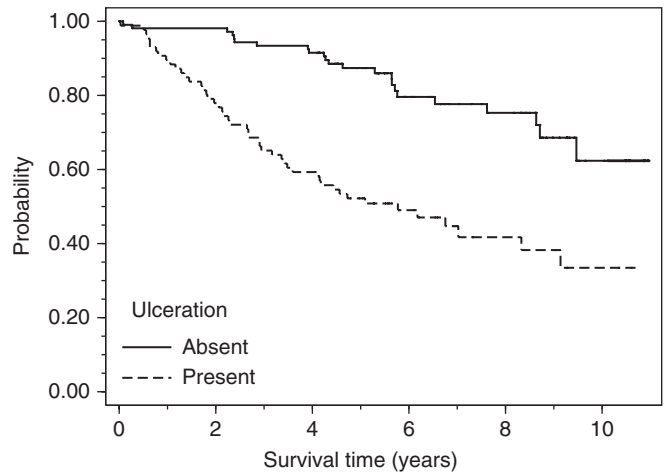

Figure 3. Kaplan–Meier estimates for malignant melanoma patients with and without ulceration.

The four variables were included simultaneously in a Cox regression model

$$\lambda(t, \mathbf{x}) = \lambda_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)$$

The estimates in this model are presented in Table 2. The top panel shows the estimated hazard ratios, obtained as $\exp(\hat{\beta}_j)$ and the corresponding 95% confidence intervals. A p-value for the test of the hypothesis of no association between the covariate and survival, that is, $\beta_j = 0$, is also shown. The estimates in Table 2 are mutually adjusted, and a comparison of these estimates with those obtained from separate analyses of each variable shows that the adjusted hazard ratio for males is considerably smaller and no longer statistically significant. This reflects that the simple analysis does not adjust for the fact that male patients have thicker tumors than females. The excess risk associated with ulceration is also reduced considerably but still statistically significant.

Cox regression analyses of the cause-specific hazards, either for death from the disease or for death from other causes, provide further insight in the association between the four predictors and the mortality

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat02177.pub2

Table 1. Single-factor logrank analyses of total mortality.

Variable	Number of patients	Deaths		logrank test (*p* value)
		Observed	Expected	
Sex of patient				
Female	126	35	46.27	7.90
Male	79	36	24.73	(0.005)
Ulceration				
Absent	115	23	44.46	27.87
Present	90	48	26.54	(<0.0001)

Table 2. Hazard ratios (HRs) with 95% confidence intervals estimated by Cox regression analysis of total mortality (top), mortality from disease (center), and mortality from other causes (bottom).

	All causes		
Variable	HR	95% CI	*p* value
Male	1.45	0.90;2.33	0.123
Age at surgery (per 10 years)	1.24	1.06;1.44	0.006
ln (thickness in *mm*)	1.54	1.13;2.09	0.006
Ulceration	2.23	1.28;3.87	0.004
	Malignant melanoma		
Variable	HR	95% CI	*p* value
Male	1.44	0.85;2.44	0.179
Age at surgery (per 10 years)	1.12	0.95;1.32	0.164
ln (thickness in *mm*)	1.74	1.23;2.48	0.002
Ulceration	2.57	1.37;4.83	0.003
	Other causes		
Variable	HR	95% CI	*p* value
Male	1.41	0.48;4.15	0.529
Age at surgery (per 10 years)	2.12	1.37;3.27	0.001
ln (thickness in *mm*)	0.96	0.52;1.80	0.908
Ulceration	1.28	0.37;4.36	0.697

of the patients. The center panel of Table 2 presents the result of a Cox regression analysis of mortality from malignant melanoma. Compared to the hazard ratios obtained from the analysis of total mortality (top panel), the estimates for tumor thickness and ulceration are increased, whereas the estimate for sex is essentially unchanged and the estimate for age at surgery is reduced and no longer statistically significant. The bottom panel of Table 2 presents the results of a Cox regression analysis of the mortality from other causes. These results should be cautiously interpreted, as the analysis is based on only 14 deaths. In general, it is not advisable to study the simultaneous effect of four variables based on so little data. In the analysis of mortality from other causes, the disease-specific covariates (thickness and ulceration) seem not important and the hazard ratio for sex is similar to the previous estimates. Age at surgery, on the other hand, is significantly associated with mortality from other causes.

The estimates from the analysis of mortality from other causes are asymptotically independent of the estimates from the analysis of mortality from malignant melanoma. A Wald test based on the difference between the estimated regression coefficients can be used to compare the effect of a given covariate on each cause of death. For age at surgery, the hazard ratios from the two cause-specific analyses are significantly different ($p = 0.007$), but for the other three covariates, this is not the case. A final word of caution in

DOI: 10.1002/9781118445112.stat02177.pub2

connection with the analyses displayed in Table 2 is that a proportional hazards model for the all cause mortality is typically mathematically incompatible with proportional cause-specific hazards models although formal goodness of fit examinations may not reject either model.

References

[1] Kalbfleisch, J.D. and Prentice, R.L. (2002) *The Statistical Analysis of Failure Time Data*, 2nd edn, John Wiley & Sons, Inc., New York.
[2] Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993) *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.
[3] Kaplan, E.L. and Meier, P. (1958) *J. Am. Stat. Assoc.*, **53**, 457–481.
[4] Aalen, O.O. (1978) *Ann. Stat.*, **6**, 701–726.
[5] Lawless, J.F. (2002) *Statistical Models and Methods for Lifetime Data*, 2nd edn, John Wiley & Sons, Inc., New York.
[6] Cox, D.R. (1972) *J. R. Stat. Soc., Ser. B*, **34**, 187–220.
[7] Martinussen, T. and Scheike, T.H. (2006) *Dynamic Regression Models for Survival Data*, Springer-Verlag, New York.
[8] Rizopoulos, D. (2012) *Joint Models for Longitudinal and Time-to-Event Data. With Applications in R.*, CRC, Boca Raton, FL.
[9] van Houwelingen, H.C. and Putter, H. (2012) *Dynamic Prediction in Clinical Survival Analysis*, CRC, Boca Raton, FL.
[10] Royston, P. and Lambert, P. (2011) *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*, Stata Press, College Station, TX.
[11] Aalen, O.O., Borgan, O., and Gjessing, H.K. (2008) *Survival and Event History Analysis. A Process Point of View*, Springer-Verlag, New York.
[12] Sun, J. (2006) *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer-Verlag, New York.
[13] Klein, J.P., van Houwelingen, H.C., Ibrahim, J.G., and Scheike, T.H. (2014) *Handbook of Survival Analysis*, CRC, Boca Raton, FL.
[14] Hougaard, P. (2000) *Analysis of Multivariate Survival Data*, Springer-Verlag, New York.
[15] Cook, R.J. and Lawless, J.F. (2007) *The Statistical Analysis of Recurrent Events*, Springer-Verlag, New York.

Wiley StatsRef: Statistics Reference Online, © 2014–2015 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat02177.pub2