Survival Analysis, Overview

Survival analysis is the study of the distribution of life times, that is, the times from an initiating event (birth, start of treatment, employment in a given job) to some terminal event (death, relapse, disability pension). A distinguishing feature of survival data is the inevitable presence of incomplete observations, particularly when the terminal event for some individuals is not observed; instead, it is only known that this event is at least later than a given point in time: *right censoring* (*see* **Censored Data**).

The aims of this entry are to provide a brief historical sketch of the long development of survival analysis and to survey what we have found to be central issues in the current methodology of survival analysis. Necessarily, this entry is rich in cross-references to other entries that treat specific subjects in more detail. However, we have not attempted to include cross-references to *all* specific entries within survival analysis.

History

The Prehistory of Survival Analysis in Demography and Actuarial Science

Survival analysis is one of the oldest statistical disciplines with roots in **demography** and **actuarial** science in the seventeenth century; see [49, Chapter 2]; [51] for general accounts of the history of **vital statistics** and [22] for specific accounts of the work before 1750.

The basic **life-table** methodology in modern terminology amounts to the estimation of a survival function (one minus distribution function) from life times with **delayed entry** (or left **truncation**; see below) and right **censoring**. This was known before 1700, and explicit **parametric models** at least since the linear approximation of de Moivre [39] (see e.g. [22, p. 517]), later examples being due to Lambert [33, p. 483]:

$$\left(1-\frac{x}{96}\right)^2 - 0.6176\left(\exp\left(-\frac{x}{31.682}\right) - \exp\left(-\frac{x}{2.43114}\right)\right) \qquad (1)$$

and the influential nineteenth-century proposals by Gompertz [19] and Makeham [37], who modeled the **hazard** function as bc^x and $a + bc^x$, respectively.

Motivated by the controversy over smallpox inoculation, D. Bernoulli [5] laid the foundation of the theory of **competing risks**; see [44] for a historical account. The calculation of **expected number of deaths** (how many deaths would there have been in a study population if a given standard set of death rates applied) also dates back to the eighteenth century; see [29] and the article on **Historical Controls in Survival Analysis**.

Among the important methodological advances in the nineteenth century was, in addition to the parametric survival analysis models mentioned above, the graphical simultaneous handling of calendar time and age in the **Lexis Diagram** [35, cf. 30].

Two very important themes of modern survival analysis may be traced to early twentieth century actuarial mathematics:

Multistate modeling in the particular case of disability insurance [41] and nonparametric estimation in continuous time of the survival function in the competing risk problem under delayed entry and right censoring [13].

At this time, survival analysis was not an integrated component of theoretical statistics. A characteristic scepticism about "the value of life-tables in statistical research" was voiced by Greenwood [20] in the *Journal of the Royal Statistical Society*, and Westergaard's [50] guest appearance in *Biometrika* on "Modern problems in vital statistics" had no reference to sampling variability. This despite the fact that these two authors were actually statistical pioneers in survival analysis: Westergaard [48] by deriving what we would call the standard error of the standardized mortality ratio (rederived by Yule [52]; see [29]) (*see* **Standardization Methods**); and Greenwood [21] with his famous expression for "the 'errors of sampling' of the survivorship tables", (see below).

The "Actuarial" life table and the Kaplan–Meier Estimator

In the mid-twentieth century, these well-established demographic and actuarial techniques were presented to the medical–statistical community in influential surveys such as those by Berkson and Gage [4] and Cutler and Ederer [13]. In this approach, time

is grouped into discrete units (e.g. one-year intervals), and the chain of survival frequencies from one interval to the next are multiplied together to form an estimate of the survival probability across several time periods. The difficulty is in the development of the necessary approximations due to the discrete grouping of the intrinsically continuous time and the possibly somewhat oblique observation fields in cohort studies and more complicated demographic situations. The penetrating study by Kaplan and Meier [28] (*see* **Kaplan–Meier Estimator**), the fascinating genesis of which was chronicled by Breslow [8], in principle, eliminated the need for these approximations in the common situation in medical statistics where all survival and censoring times are known precisely. Kaplan and Meier's tool (which they traced back to Böhmer [7]) was to shrink the observation intervals to include at most one observation per interval. Though overlooked by many later authors, Kaplan and Meier also formalized the age-old handling of **delayed entry** (actually also covered by Böhmer) through the necessary adjustment for the **risk set**, the set of individuals alive and under observation at a particular value of the relevant time variable.

Among the variations on the actuarial model, we will mention two.

Harris et al. [23] anticipated much recent work in, for example, **AIDS** survival studies in their generalization of the usual life-table estimator to the situation in which the death and censoring times are known only in large, irregular intervals (*see* **Grouped Survival Times**).

Ederer et al. [(14)] developed a "relative survival rate... as the ratio of the observed survival rate in a group of patients to the survival rate expected in a group similar to the patients ..." thereby connecting to the long tradition of comparing observed with expected; see, for example, [29] and the article on **Historical Controls in Survival Analysis**.

Parametric Survival Models

Parametric survival models were well-established in actuarial science and demography, but have never dominated medical uses of survival analysis. However, in the 1950s and 1960s important contributions to the statistical theory of survival analysis were based on simple parametric models. One example is the **maximum likelihood** approach by Boag [6] to a **cure model** assuming eternal life with probability c and **lognormally distributed** survival times otherwise. The **exponential distribution** was assumed by Littell [36], when he compared the "actuarial" and the maximum likelihood approach to the "T-year survival rate", by Armitage [3] in his comparative study of two-sample tests for **clinical trials** with **staggered entry**, and by Feigl and Zelen [16] in their model for (uncensored) lifetimes whose expectations were allowed to depend linearly on covariates, generalized to censored data by Zippin and Armitage [53].

Cox [11] revolutionized survival analysis by his **semiparametric regression** model for the hazard, depending arbitrarily ("nonparametrically") on time and parametrically on covariates (*see* **Cox Regression Model**). For details on the genesis of Cox's paper, see [42, 43].

Multistate Models

Traditional actuarial and demographical ways of modeling several life events simultaneously may be formalized within the probabilistic area of finite-state **Markov processes** in continuous time. An important and influential documentation of this was by Fix and Neyman [18], who studied recovery, relapse, and death (and censoring) in what is now commonly termed an illness–death model allowing for competing risks (*see* **Fix–Neyman Process**). Chiang [9], for example, in his 1968 monograph, extensively documented the relevant stochastic models (*see* **Stochastic Processes**), and Sverdrup [46], in an important paper, gave a systematic statistical study. These models have constant transition intensities, although subdivision of time into intervals allows grouped-time methodology of the actuarial life-table type, as carefully documented by Hoem [24].

Survival Analysis Concepts

The ideal basic independent nonnegative **random variables** $X_i, i = 1, \ldots, n$ are not always observed directly. For some individuals i, the available piece of information is a *right-censoring* time U_i, that is, a period elapsed in which the event of interest has not occurred (e.g. a patient has survived until U_i). Thus, a generic survival data sample includes $((\widetilde{X}_i, D_i), i = 1, \ldots, n)$ where $\widetilde{X}_i$ is the smaller of X_i and U_i and D_i is the indicator, $I(X_i \leq U_i)$, of not being censored.

Mathematically, the distribution of X_i may be described by the *survival function*

$$S_i(t) = \Pr(X_i > t). \tag{2}$$

If the **hazard function**

$$\alpha_i(t) = \lim_{\Delta t \to 0} \frac{\Pr(X_i \le t + \Delta t \mid X_i > t)}{\Delta t} \tag{3}$$

exists, then

$$S_i(t) = \exp(-A_i(t)), \tag{4}$$

where

$$A_i(t) = \int_0^t \alpha_i(u)\,du \tag{5}$$

is the integrated hazard over $[0, t)$. If, more generally, the distribution of the X_i has discrete components, then $S_i(t)$ is given by the **product-integral** of the cumulative hazard measure. Owing to the dynamical nature of survival data, a characterization of the distribution via the hazard function is often convenient. (Note that $\alpha_i(t)\Delta t$ when $\Delta t > 0$ is *small* is approximately the conditional probability of i "dying" just after time t given "survival" till time t.) Also, $\alpha_i(t)$ is the basic quantity in the **counting process** approach to survival analysis (see e.g. [2], and the article on **Survival Distributions and Their Characteristics**).

Nonparametric Estimation and Testing

The simplest situation encountered in survival analysis is the nonparametric estimation of a survival distribution function based on a right-censored sample of observation times $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$, where the true survival times $X_i, i = 1, \ldots, n$, are assumed to be independent and identically distributed with common survival distribution function $S(t)$, whereas as few assumptions as possible are usually made about the right-censoring times U_i except for the assumption of *independent censoring* (*see* **Censored Data**). The concept of independent censoring has the interpretation that the fact that an individual, i, is alive *and uncensored* at time t, say, should not provide more information on the survival time for that individual than $X_i > t$, that is, the right-censoring mechanism should not remove individuals from the study who are at a particularly high or a particularly low risk of dying. Under these assumptions, $S(t)$ is estimated by the *Kaplan–Meier estimator* [28]. This is given by

$$\widehat{S(t)} = \prod_{\widetilde{X}_i \le t} \left[1 - \frac{D_i}{Y(\widetilde{X}_i)}\right], \tag{6}$$

where $Y(t) = \sum I(\widetilde{X}_i \ge t)$ is the number of individuals *at risk* just before time t. The Kaplan–Meier estimator is a **nonparametric maximum likelihood** estimator and, in large samples, $\widehat{S(t)}$ is approximately normally distributed with mean $S(t)$ and a variance that may be estimated by Greenwood's formula:

$$\widehat{\sigma^2(t)} = [\widehat{S(t)}]^2 \sum_{\widetilde{X}_i \le t} \frac{D_i}{Y(\widetilde{X}_i)[Y(\widetilde{X}_i) - 1]}. \tag{7}$$

From this result, pointwise **confidence intervals** for $S(t)$ are easily constructed and, since one can also show weak **convergence** of the entire Kaplan–Meier curve $\{\sqrt{(n)}[\widehat{S(t)} - S(t)]; 0 \le t \le \tau\}, \tau \le \infty$ to a mean zero Gaussian process (*see* **Brownian Motion and Diffusion Processes**), simultaneous confidence bands for $S(t)$ on $[0, \tau]$ can also be set up.

As an alternative to estimating the survival distribution function $S(t)$, the *cumulative hazard function* $A(t) = -\log S(t)$ may be studied. Thus, $A(t)$ may be estimated by the **Nelson–Aalen Estimator**

$$\widehat{A(t)} = \sum_{\widetilde{X}_i \le t} \frac{D_i}{Y(\widetilde{X}_i)}. \tag{8}$$

The relation between the estimators $\widehat{S(t)}$ and $\widehat{A(t)}$ is given by the *product-integral* from which it follows that their large-sample properties are equivalent. Though the Kaplan–Meier estimator has the advantage that a survival *probability* is easier to interpret than a cumulative hazard function, the Nelson–Aalen estimator is easier to generalize to multistate situations beyond the survival data context. We shall return to this below. To give a nonparametric estimate of the hazard function $\alpha(t)$ itself requires some smoothing techniques to be applied (see **Smoothing Hazard Rates**).

Right censoring is not the only kind of data-incompleteness to be dealt with in survival analysis; in particular, *left truncation* (or **delayed entry**) where individuals may not all be followed from time 0 but maybe from a later entry time V_i conditionally on having survived until V_i, occurs frequently in, for example, epidemiological applications. Dealing with

left truncation only requires a redefinition of the *risk set* from the set $\{i: \widetilde{X}_i \geq t\}$ of individuals still alive and uncensored at time t to the set $\{i: V_i < t \leq \widetilde{X}_i\}$ of individuals with entry time $V_i < t$ and who are still alive and uncensored. With $Y(t)$ still denoting the size of the risk set at time t both (6), (7), and (8) are applicable though one should be aware of the fact that estimates of $S(t)$ and $A(t)$ may be ill-determined for small values of t due to the left truncation (*see* **Truncated Survival Times**).

When the survival time distributions in a number, k, of homogeneous groups have been estimated nonparametrically, it is often of interest to test the hypothesis H_0 of identical hazards in all groups. Thus, on the basis of censored survival data $((\widetilde{X}_{hi}, D_{hi}),\ i = 1, \ldots, n_h)$ for group $h, h = 1, \ldots, k$, the Nelson–Aalen estimates $\widehat{A_h(t)}$ have been computed, and based on the combined sample of size $n = \sum_h n_h$ with data $((\widetilde{X}_i, D_i), i = 1, \ldots, n)$, an estimate of the common cumulative hazard function $A(t)$ under H_0 may be obtained by a Nelson–Aalen estimator $\widehat{A(t)}$. As a general statistic for testing H_0, one may then use a k-vector of sums of weighted differences between the increments of $\widehat{A_h(t)}$ and $\widehat{A(t)}$:

$$Z_h = \sum_{i=1}^{n} K_h(\widetilde{X}_i)[\Delta\widehat{A_h}(\widetilde{X}_i) - \Delta\hat{A}(\widetilde{X}_i)]. \quad (9)$$

Here, $\Delta\widehat{A_h(t)} = 0$ if t is not among the observed survival times in the hth sample and $K_h(t)$ is 0 whenever $Y_h(t) = 0$, in fact all weight functions used in practice have the form $K_h(t) = Y_h(t)K(t)$. With this structure for the weight function, the covariance between Z_h and Z_j given by (9) is estimated by

$$\sigma_{hj} = \sum_{i=1}^{n} K^2(\widetilde{X}_i)\frac{Y_h(\widetilde{X}_i)}{Y(\widetilde{X}_i)}\left[\delta_{hj} - \frac{Y_j(\widetilde{X}_i)}{Y(\widetilde{X}_i)}\right]D_i, \quad (10)$$

and, letting $\mathbf{Z}$ be the k-vector $(Z_1, \ldots, Z_k)'$ and $\boldsymbol{\Sigma}$ the k by k matrix $(\sigma_{hj}, h, j = 1, \ldots, k)$ the test statistic $X^2 = \mathbf{Z}'\boldsymbol{\Sigma}^{-}\mathbf{Z}$ is asymptotically **chi-squared distributed** under H_0 with $k - 1$ **degrees of freedom** if all n_h tend to infinity at the same rate. Here, $\boldsymbol{\Sigma}^{-}$ is a generalized inverse for $\boldsymbol{\Sigma}$ (*see* **Matrix Algebra**).

Special choices for $K(t)$ correspond to test statistics with different properties for particular alternatives to H_0 (*see* **Linear Rank Tests in Survival Analysis**). An important such test statistic is the **logrank** test obtained for $K(t) = I(Y(t) > 0)$. For this test, which has particularly good power for **proportional hazards** alternatives, Z_h given by (9) reduces to $Z_h = O_h - E_h$ with O_h the total number of observed failures in group h and $E_h = \sum D_i Y_h(\widetilde{X}_i)/Y(\widetilde{X}_i)$ an "expected" number of failures in group h. For the two-sample case $(k = 2)$, one may of course use the square root of X^2 as an asymptotically normal test statistic for the **null hypothesis**. For the case where the k groups are *ordered*, and where a *score* x_h (with $x_1 \leq \ldots \leq x_k$) is attached to group h, a *test for trend* is given by $T^2 = (\mathbf{x}'\mathbf{Z})^2/\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}$ with $\mathbf{x} = (x_1, \ldots, x_k)'$ and it is asymptotically chi-squared with 1 df.

The above **linear rank tests** have low **power** against certain important classes of alternatives such as "crossing hazards". Just as for uncensored data, this has motivated the development of test statistics of the **Kolmogorov–Smirnov and Cramèr–von Mises** types, based on maximal deviation or integrated squared deviation between estimated hazards, cumulative hazards or survival functions.

Parametric Inference

The nonparametric methods outlined in the previous section have become the standard approach to the analysis of simple homogeneous survival data without covariate information. However, parametric survival time distributions are sometimes used for inference, and we shall here give a brief review. Assume again that the true survival times $X_1, \ldots, X_n$ are independent and identically distributed with survival distribution function $S(t; \boldsymbol{\theta})$ and hazard function $\alpha(t; \boldsymbol{\theta})$ but that only a right-censored sample $(\widetilde{X}_i, D_i), i = 1, \ldots, n$, is observed. Under independent censoring, the likelihood function for the parameter $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}(\alpha(\widetilde{X}_i; \boldsymbol{\theta}))^{D_i} S(\widetilde{X}_i; \boldsymbol{\theta}). \quad (11)$$

The function (11) may be analyzed using standard **large-sample theory**. Thus, standard tests, that is, Wald-, score-, and **likelihood ratio tests** are used as inferential tools (*see* **Chi-square Tests**). Two frequently used parametric survival models are the **Weibull distribution** with hazard function

$\alpha\rho(\alpha t)^{\rho-1}$, and the piecewise exponential distribution with $\alpha(t, \boldsymbol{\theta}) = \alpha_j$ for $t \in I_j$ with $I_j = [t_{j-1}, t_j)$, $0 = t_0 < t_j < \cdots < t_m = \infty$. Both of these distributions contain the very simplest model, the **exponential distribution** with a constant hazard function as null cases (*see* **Parametric Models in Survival Analysis**).

Comparison with Expected Survival

As a special case of the nonparametric tests discussed above, a *one-sample* situation may be studied. This may be relevant if one wants to compare the observed survival in the sample with the *expected survival* based on a standard life table. Thus, assume that a hazard function $\alpha^*(t)$ is given and that the hypothesis $H_0 : \alpha = \alpha^*$ is to be tested. One test statistic for H_0 is the one-sample *logrank test* $(O - E^*)/(E^*)^{1/2}$ where E^*, the "expected" number of deaths is given by $E^* = \sum[A^*(\widetilde{X}_i) - A^*(V_i)]$ (with A^* the cumulative hazard corresponding to α^*). In this case, $\widehat{\theta} = O/E^*$, the *standardized mortality ratio*, is the maximum likelihood estimate for the parameter θ in the model $\alpha(t) = \theta\alpha^*(t)$. Thus, the standardized mortality ratio arises from a **multiplicative model** involving the known population hazard $\alpha^*(t)$. Another classical tool for comparing with expected survival, the so-called *expected survival function*, arises from an *additive or excess* hazard model (*see* **Excess Mortality**; **Expected Number of Deaths**; **Historical Controls in Survival Analysis**).

The Cox Regression Model

In many applications of survival analysis, the interest focuses on how *covariates* may affect the outcome; in clinical trials, adjustment of treatment effects for effects of other **explanatory variables** may be crucial if the randomized groups are unbalanced with respect to important **prognostic factors**, and in epidemiological **cohort studies**, reliable effects of exposure may be obtained only if some adjustment is made for **confounding** variables. In these situations, a *regression model* is useful and the most important model for survival data is the **Cox** [11] *proportional hazards regression model*. In its simplest form, it states the hazard function for an individual, i, with covariates $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})'$ to be

$$\alpha_i(t; \boldsymbol{Z}_i) = \alpha_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i), \qquad (12)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown regression coefficients and $\alpha_0(t)$, the *baseline hazard*, is the hazard function for individuals with all covariates equal to 0. Thus, the baseline hazard describes the common shape of the survival time distributions for all individuals while the *relative risk* function $\exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)$ gives the level of each individual's hazard. The interpretation of the parameter, β_j for a dichotomous $Z_{ij} \in \{0, 1\}$ is that $\exp(\beta_j)$ is the **relative risk** for individuals with $Z_{ij} = 1$ compared to those with $Z_{ij} = 0$ all other covariates being the same for the two individuals. Similar interpretations hold for parameters corresponding to covariates taking more than two values.

The model is **semiparametric** in the sense that the relative risk part is modeled parametrically while the baseline hazard is left unspecified. This semiparametric nature of the model led to a number of inference problems, which was discussed in the literature in the years following the publication of Cox's article in 1972. However, these problems were all resolved and estimation proceeds as follows. The regression coefficients $\boldsymbol{\beta}$ are estimated by maximizing the **Cox partial likelihood**

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{Z}_j)} \right]^{D_i}, \qquad (13)$$

where $R_i = \{j : \widetilde{X}_j \geq \widetilde{X}_i\}$, the **risk set** at time $\widetilde{X}_i$, is the set of individuals still alive and uncensored at that time. Furthermore, the cumulative baseline hazard $A_0(t)$ is estimated by the *Breslow estimator*

$$\widehat{A_0(t)} = \sum_{\widetilde{X}_i \leq t} \frac{D_i}{\sum_{j \in R_i} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{Z}_j)}, \qquad (14)$$

which is the Nelson–Aalen estimator one would use if $\boldsymbol{\beta}$ were known and equal to the maximum partial likelihood estimate $\widehat{\boldsymbol{\beta}}$. The estimators based on (13) and (14) also have a nonparametric maximum likelihood interpretation. In large samples, $\widehat{\boldsymbol{\beta}}$ is approximately normally distributed with the proper mean and with a **covariance**, which is estimated by the **information matrix** based on (13). This means that approximate confidence intervals for the relative risk parameters of interest can be calculated and that the usual large-sample test statistics based on (13) are available. Also, the asymptotic distribution of the Breslow estimator is normal; however, this estimate

is most often used as a tool for estimating *survival probabilities* for individuals with given covariates, $\mathbf{Z}_0$. Such an estimate may be obtained by the product integral $\widehat{S(t;\mathbf{Z}_0)}$ of $\exp(\widehat{\boldsymbol{\beta}}'\mathbf{Z}_0)\widehat{A_0(t)}$. The *joint* asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and the Breslow estimator then yields an approximate normal distribution for $\widehat{S(t;\mathbf{Z}_0)}$ in large samples.

A number of useful extensions of this simple Cox model are available. Thus, in some cases, the covariates are **time-dependent**, for example, a covariate might indicate whether or not a given event had occurred by time t, or a time-dependent covariate might consist of repeated recordings of some measurement likely to affect the prognosis. In such cases, the regression coefficients $\boldsymbol{\beta}$ are estimated replacing $\exp(\boldsymbol{\beta}'\mathbf{Z}_j)$ in (13) by $\exp[\boldsymbol{\beta}'\mathbf{Z}_j(\widetilde{X}_i)]$.

Also, a simple extension of the Breslow estimator (14) applies in this case. However, the survival function can, in general, no longer be estimated in a simple way because of the extra randomness arising from the covariates, which is not modeled in the Cox model. This has the consequence that the estimates are more difficult to interpret when the model contains time-dependent covariates. To estimate the survival function in such cases, a joint model for the hazard and the time-dependent covariate is needed (*see* **Joint Modeling of Longitudinal and Event Time Data**).

Another extension of (12) is the *stratified* Cox model where individuals are grouped into a number, k of strata each of which has a separate baseline hazard (*see* **Stratification**). This model has important applications for checking the assumptions of (12). The model assumption of proportional hazards may also be *tested* in a number of ways, the simplest possibility being to add interaction terms of the form $Z_{ij}f(t)$ between Z_{ij} and time where $f(t)$ is some specified function. Also, various forms of *residuals* as for normal linear models may be used for **model checking** in (12) (*see* **Goodness of Fit in Survival Analysis**; **Residuals for Survival Analysis**). In (12), it is finally assumed that a quantitative covariate affects the hazard *log-linearly*. This assumption may also be checked in several ways and alternative models with other relative risk functions $r(\boldsymbol{\beta}'\mathbf{Z}_i)$ may be used. Special care is needed when covariates are measured with error (*see* **Measurement Error in Survival Analysis**).

Other Regression Models for Survival Data

Though the semiparametric Cox model is the regression model for survival data that is applied most frequently, other regression models, for example, *parametric* regression models also play important roles in practice. Examples include models with a multiplicative structure, that is, models like (12) but with a parametric specification, $\alpha_0(t) = \alpha_0(t;\boldsymbol{\theta})$, of the baseline hazard, and **accelerated failure-time models**.

A multiplicative model with important epidemiological applications is the **Poisson regression** model with a piecewise constant baseline hazard. In large data sets with categorical covariates, this model has the advantage that a sufficiency reduction to the number of failures and the amount of person-time at risk in each *cell* defined by the covariates and the division of time into intervals is possible. This is in contrast to the Cox regression model (12) where each individual data record is needed to compute (13). The substantial computing time required to maximize (13) in large samples has also led to modifications of this estimation procedure. Thus, in *nested case–control studies* the risk set R_i in the Cox partial likelihood is replaced by a random sample $\widetilde{R}_i$ of R_i (*see* **Case–Control Study, Nested**).

In the accelerated failure-time model, the focus is not on the hazard function but on the survival time itself much like in classical linear models. Thus, this model is given by $\log X_i = \alpha + \boldsymbol{\beta}'\mathbf{Z}_i + \varepsilon_i$, where the error terms are assumed to be independent and identically distributed with expectation 0. Examples include **normally distributed** $(\varepsilon_i, i = 1, \ldots, n)$, and error terms with a **logistic** or an **extreme value** distribution, the latter giving rise to a regression model with **Weibull** distributed life times.

Finally, we shall mention some nonparametric hazard models. In Aalen's additive model, $\alpha_i(t) = \beta_0(t) + \boldsymbol{\beta}(t)'\mathbf{Z}_i(t)$ (*see* **Aalen's Additive Regression Model**), the regression functions $\beta_0(t), \ldots, \beta_p(t)$ are left completely unspecified and estimated nonparametrically much like the Nelson–Aalen estimator discussed above. This model provides an attractive alternative to the other regression models discussed in this section. There also exist more general and flexible models containing both this model and the Cox regression model as special cases (*see* **Additive–Multiplicative Intensity Models**).

Multistate Models

Models for survival data may be considered a special case of a *multistate model*, namely, a model with a transient state *alive* (0) and an absorbing state *dead* (1) and where the hazard rate is the force of transition from state 0 to state 1. Multistate models may conveniently be studied in the mathematical framework of *counting processes* with a notation that actually simplifies the notation of the previous sections and, furthermore, unifies the description of survival data and that of more general models like the competing risks model and the illness–death model to be discussed below. We first introduce the counting processes relevant for the study of censored survival data [1] Define, for $i = 1, \ldots, n$, the stochastic processes

$$N_i(t) = I(\widetilde{X}_i \leq t, D_i = 1) \qquad (15)$$

and

$$Y_i(t) = I(\widetilde{X}_i \geq t). \qquad (16)$$

Then (15) is a counting process counting 1 at time $\widetilde{X}_i$ if individual i is observed to die; otherwise $N_i(t) = 0$ throughout. The process (16) indicates whether i is still at risk just before time t. Models for the survival data are then introduced via the *intensity process*, $\lambda_i(t) = \alpha_i(t)Y_i(t)$ for $N_i(t)$, where $\alpha_i(t)$, as before, denotes the hazard function for the distribution of X_i. Letting $N = N_1 + \cdots + N_n$ and $Y = Y_1 + \cdots + Y_n$ the Nelson–Aalen estimator (8) is given by the stochastic integral

$$\widehat{A(t)} = \int_0^t \frac{J(u)}{Y(u)}\, \mathrm{d}N(u), \qquad (17)$$

where $J(t) = I(Y(t) > 0)$. In this simple multistate model, the *transition probability* $P_{00}(0, t)$, that is, the conditional probability of being in state 0 by time t given state 0 at time 0 is simply the survival probability $S(t)$, which, as described above, may be estimated using the Kaplan–Meier estimator, which is the product-integral of (17). In fact, all the models and methods for survival data discussed above, which are based on the hazard function have immediate generalizations to models based on counting processes. Thus, both the nonparametric tests and the Cox regression model may be applied for counting process (multistate) models (*see* **Counting Process Methods in Survival Analysis**).

One important extension of the two-state model for survival data is the *competing risks* model with one transient alive state 0 and a number, k, of absorbing states corresponding to death from cause h, $h = 1, \ldots, k$. In this model, the basic parameters are the cause-specific hazard functions $\alpha_h(t)$, $h = 1, \ldots, k$, and the observations for individual i will consist of $(\widetilde{X}_i, D_{hi})$, $h = 1, \ldots, k$, where $D_{hi} = 1$ if individual i is observed to die from cause h, and $D_{hi} = 0$ otherwise. On the basis of these data, k counting processes for each i can be defined by $N_{hi}(t) = I(\widetilde{X}_i \leq t, D_{hi} = 1)$ and letting $N_h = N_{h1} + \cdots + N_{hn}$, the integrated cause-specific hazard $A_h(t)$ is estimated by the Nelson–Aalen estimator replacing N by N_h in (17). A useful synthesis of the cause-specific hazards is provided by the transition probabilities $P_{0h}(0, t)$ of being dead from cause h by time t. This is frequently called the *cumulative incidence function* for cause h and is given by

$$P_{0h}(s, t) = \int_s^t S(u)\alpha_h(u)\, \mathrm{d}u, \qquad (18)$$

and hence it may be estimated by (18) by inserting the Kaplan–Meier estimate for $S(u)$ and the Nelson–Aalen estimate for the integrated cause-specific hazard. In fact, this **Aalen–Johansen estimator** of the matrix of transition probabilities is exactly the product-integral of the cause-specific hazards.

Another important multistate model is the *illness–death* or *disability* model with two transient states, say *healthy* (0) and *diseased* (1) and one absorbing state *dead* (2). If transitions both from 0 to 1 and from 1 to 0 are possible, the disease is *recurrent*, otherwise it is *chronic*. On the basis of such observed transitions between the three states, it is possible to define counting processes for individual i as $N_{hj\,i}(t) =$ number of observed $h \to j$ transitions in the time interval $[0, t]$ for individual i and, furthermore, we may let $Y_{hi}(t) = I$ (i is in state h at time $t-$). With these definitions, we may set up and analyze models for the transition intensities $\alpha_{hj\,i}(t)$ from state h to state j including nonparametric comparisons and Cox-type regression models. Furthermore, transition probabilities $P_{hj}(s, t)$ may be estimated by product-integration of the intensities.

Other Kinds of Incomplete Observation

A salient feature of survival data is *right censoring*, which has been referred to throughout in the present

overview. However, several other kinds of incomplete observation are important in survival analysis.

Often, particularly when the time variable of interest is age, individuals enter study after time 0. This is called *delayed entry* and may be handled by *left truncation* (conditioning) or *left filtering* ("viewing the observations through a filter"). There are also situations when only events (such as AIDS cases) that occur *before* a certain time are included (*right truncation*) (*see* **Truncated Survival Times**). The phenomenon of *left censoring*, though theoretically possible, is more rarely relevant in survival analysis.

When the event times are only known to lie in an interval, one may use the *grouped time* approach of classical *life tables* (*see* **Grouped Survival Times**; **Life Table**), or (if the intervals are not synchronous) techniques for **interval censoring** may be relevant.

A common framework (**coarsening at random**) was recently suggested for several of the above types of incomplete observation.

Multivariate Survival Analysis

For **multivariate survival**, the innocently looking problem of generalizing the Kaplan–Meier estimator to several dimensions has proved surprisingly intricate. A major challenge (in two dimensions) is how to efficiently use singly censored observations, where one component is observed and the other is right censored.

For regression analysis of multivariate survival times, two major approaches have been taken. One is to model the marginal distributions and use estimation techniques based on **generalized estimating equations** leaving the association structure unspecified (see **Marginal Models for Multivariate Survival Data**.) The other is to specify **random effects** models for survival data based on conditional independence (*see* **Frailty**.) An interesting combination between these two methods is provided by **copula** models in which the marginal distributions are combined via a so-called copula function thereby obtaining an explicit model for the association structure.

For the special case of repeated events, both the marginal approach and the conditional (frailty) approach have been used successfully (*see* **Repeated Events**).

Concluding Remarks

Survival analysis is a well-established discipline in statistical theory as well as in biostatistics. Most books on biostatistics contain chapters on the topic and most **software** packages include procedures for handling the basic survival techniques (*see* **Survival Analysis, Software**). Several books have appeared, among them the documentation of the actuarial and demographical know-how by Elandt–Johnson and Johnson [15]; the research monograph by Kalbfleisch and Prentice [27], the first edition of which for a decade maintained its position as main reference on the central theory; the comprehensive text by Lawless [34] covering also parametric models, and the concise text by Cox and Oakes [12], two central contributors to the recent theory. The counting process approach is covered by Fleming and Harrington [17] and by Andersen et al. [2]; see also [25]. Later, books intended primarily for the biostatistical user have appeared. These include [10, 31, 32, 38, 40]. Also, books dealing with special topics, like implementation in the **S-Plus** software [47], multivariate survival data [26], and the linear regression model [45] have appeared.

References

[1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals. of Statistics* **6**, 701–726.

[2] Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

[3] Armitage, P. (1959). The comparison of survival curves, *Journal of the Royal Statistical Society A* **122**, 279–300.

[4] Berkson, J. & Gage, R.P. (1950). Calculation of survival rates for cancer, *Proceedings of the Staff Meetings of the Mayo Clinic* **25**, 270–286.

[5] Bernoulli, D. (1766). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir, *Mémoires de Mathématique et de Physique de l'Académie Royale des Sciences, Paris* Année MDCCLX, pp. 1–45 of Mémoires.

[6] Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society B* **11**, 15–53.

[7] Böhmer, P.E. (1912). Theorie der unabhängigen Wahrscheinlichkeiten, *Rapports, Mémoires et Procés – verbaux du 7^e Congrès International d'Actuaires, Amsterdam* **2**, 327–343.

[8] Breslow, N.E. (1991). Introduction to Kaplan and Meier (1958). Nonparametric estimation from incomplete observations, in *Breakthroughs in Statistics II*, S. Kotz & N.L. Johnson, eds. Springer, New York, 311–318.

[9] Chiang, C.L. (1968). *Introduction to Stochastic Processes in Biostatistics*. Wiley, New York.

[10] Collett, D. (2003). *Modelling Survival Data in Medical Research*, 2nd Ed., Chapman and Hall, London.

[11] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society (B)* **34**, 187–220.

[12] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

[13] Cutler, S.J. & Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival, *Journal of Chronic Diseases* **8**, 699–713.

[14] Ederer, F., Axtell, L.M. & Cutler, S.J. (1961). The relative survival rate: A statistical methodology, *National Cancer Institute Monographs* **6**, 101–121.

[15] Elandt-Johnson, R.C. & Johnson, N.L. (1980). *Survival Models and Data Analysis*. Wiley, New York.

[16] Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics* **21**, 826–838.

[17] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

[18] Fix, E. & Neyman, J. (1951). A simple stochastic model of recovery, relapse, death and loss of patients, *Human Biology* **23**, 205–241.

[19] Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, *Philosophical Transactions of the Royal Society of London, Series A* **115**, 513–580.

[20] Greenwood, M. (1922). Discussion on the value of life-tables in statistical research, *Journal of the Royal Statistical Society* **85**, 537–560.

[21] Greenwood, M. (1926). The natural duration of cancer, in *Reports on Public Health and Medical Subjects*, Vol. 33 His Majesty's Stationery Office, London, pp. 1–26.

[22] Hald, A. (1990). *A History of Probability and Statistics and Their Applications before 1750*. Wiley, New York.

[23] Harris, T.E., Meier, P. & Tukey, J.W. (1950). The timing of the distribution of events between observations, *Human Biology* **22**, 249–270.

[24] Hoem, J.M. (1976). The statistical theory of demographic rates. A review of current developments (with discussion), *Scandinavian Journal of Statistics* **3**, 169–185.

[25] Hosmer, D.W. & Lemeshow, S. (1999). *Applied Survival Analysis. Regression Modeling of Time to Event Data*. Wiley, New York.

[26] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer, New York.

[27] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed., Wiley, New York.

[28] Kaplan, E.L. & Meier, P. (1958). Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481, 562–563.

[29] Keiding, N. (1987). The method of expected number of deaths 1786–1886–1986, *International Statistical Review* **55**, 1–20.

[30] Keiding, N. (1990). Statistical inference in the Lexis diagram, *Philosophical Transactions of the Royal Society London A* **332**, 487–509.

[31] Klein, J.P. & Moeschberger, M.L. (2003). *Survival Analysis. Techniques for Censored and Truncated Data*, 2nd Ed., Springer, New York.

[32] Kleinbaum, D.G. (1996). *Survival Analysis. A Self-Learning Text*. Springer, New York.

[33] Lambert, J.H. (1772). *Beyträge zum Gebrauche der Mathematik und deren Anwendung*, Vol. III, Verlage des Buchlages der Realschule, Berlin.

[34] Lawless, J.F. (2002). *Statistical Models and Methods for Lifetime Data*, 2nd Ed., Wiley, New York.

[35] Lexis, W. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. Trübner, Strassburg.

[36] Littell, A.S. (1952). Estimation of the T-year survival rate from follow-up studies over a limited period of time, *Human Biology* **24**, 87–116.

[37] Makeham, W.M. (1860). On the law of mortality, and the construction of mortality tables, *Journal of the Institute of Actuaries* **8**, 301.

[38] Marubini, E. & Valsecchi, M.G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley, Chichester.

[39] de Moivre, A. (1725). *Annuities upon Lives: or, The Valuation of Annuities uon any Number of Lives; as also, of Reversions. To which is added, An Appendix concerning the Expectations of Life, and Probabilites of Survivorship*. Fayram, Motte and Pearson, London.

[40] Parmar, K.B. & Machin, D. (1995). *Survival analysis. A practical approach*. Wiley, Chichester.

[41] du Pasquier, L.G. (1913). Mathematische Theorie der Invaliditätsversicherung, *Mitteilungen der Vereinigung der Schweizerische Versicherungs-Mathematiker* **8**, 1–153.

[42] Prentice, R.L. (1991). Introduction to Cox (1972) Regression models and life-tables, in *Breakthroughs in Statistics II*, S. Kotz & N.L. Johnson eds. Springer, New York, pp. 519–526.

[43] Reid, N. (1994). A conversation with Sir David Cox, *Statistical Science* **9**, 439–455.

[44] Seal, H.L. (1977). Studies in the history of probability and statistics, XXXV. Multiple decrements or competing risks, *Biometrika* **64**, 429–439.

[45] Smith, P.J. (2002). *Analysis of Failure Time Data*. Chapman and Hall/CRC, London.

[46] Sverdrup, E. (1965). Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health, *Skandinavisk Aktuarietidskrift* **48**, 184–211.

[47] Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.
[48] Westergaard, H. (1882). *Die Lehre von der Mortalität und Morbilität*. Fischer, Jena.
[49] Westergaard, H. (1901). *Die Lehre von der Mortalität und Morbilität*, 2. Aufl., Fischer, Jena.
[50] Westergaard, H. (1925). Modern problems in vital statistics, *Biometrika* **17**, 355–364.
[51] Westergaard, H. (1932). *Contributions to the History of Statistics*. King, London.
[52] Yule, G. (1934). On some points relating to vital statistics, more especially statistics of occupational mortality (with discussion), *Journal of the Royal Statistical Society* **97**, 1–84.
[53] Zippin, C. & Armitage, P. (1966). Use of concomitant variables and incomplete survival information with estimation of an exponential survival parameter, *Biometrics* **22**, 655–672.

PER KRAGH ANDERSEN & NIELS KEIDING