BIOMETRICS 55, 13–22
March 1999

Fundamentals of Survival Data

Philip Hougaard

Novo Nordisk, Novo Alle Building 9F2, DK-2880 Bagsvaerd, Denmark
email: pho@novo.dk

SUMMARY. Survival data stand out as a special statistical field. This paper tries to describe what survival data is and what makes it so special. Survival data concern times to some events. A key point is the successive observation of time, which on the one hand leads to some times not being observed so that all that is known is that they exceed some given times (censoring), and on the other hand implies that predictions regarding the future course should be conditional on the present status (truncation). In the simplest case, this condition is that the individual is alive. The successive conditioning makes the hazard function, which describes the probability of an event happening during a short interval given that the individual is alive today (or more generally able to experience the event), the most relevant concept. Standard distributions available (normal, log-normal, gamma, inverse Gaussian, and so forth) can account for censoring and truncation, but this is cumbersome. Besides, they fit badly because they are either symmetric or right skewed, but survival time distributions can easily be left-skewed positive variables. A few distributions satisfying these requirements are available, but often nonparametric methods are preferable as they account better conceptually for truncation and censoring and give a better fit. Finally, we compare the proportional hazards regression models with accelerated failure time models.

KEY WORDS: Accelerated failure times; Censoring; Frailty; Hazard function; Survival data; Time-dependent covariates; Truncation.

1. Introduction

First, we describe what survival data is in Section 2. Then the paper considers several fundamental questions on survival data. What makes such data different from other types of data? In other words, why does the subject need a special statistical theory? The answer is given in Section 3. As an extra result of these considerations, we get answers to two other questions: Why is the hazard function so important? Why are nonparametric methods so popular? Furthermore, we compare and discuss two alternatives for modeling the effect of explanatory variables, the proportional hazards regression models and the accelerated failure time models, in Section 4. Neglected covariates and time-dependent covariates are also discussed in Section 4.

2. What Is Survival Data?

Survival data is a term used for data measuring the time to some event. In the simplest case, the event is death, but the term also covers other events, like occurrence of a disease, a complication, or, e.g., the time to occurrence of an epileptic seizure. In industrial applications, it is typically time to failure of a unit or some component in a unit. In economics, it can be time to acceptance of a job offer for an unemployed person. In demography, the event can be entering marriage. In many cases, the event is a transition from one state to another. For example, death is a transition from the state alive to the state dead (see Figure 1). Occurrence of disease is a transition from being healthy to a state of presence of disease (see Figure 2). In the demographic example, it is a transition from single to married. For the epileptic seizure, the strict definition of an event is the transition from the seizure-free state to the state of active seizure, but for practical purposes, the duration of a seizure is short compared to the time studied and therefore we consider the whole seizure an event rather than just the start of it. To accommodate this, we define states corresponding to the accumulated number of seizures the patient has experienced.

In the figures, the name of the transition is included above the arrow. Depending on the context, we use words like death, event, failure, and transition to cover the same thing, namely what happens at the response time. In some cases, the interesting aspect is the transition, corresponding to the incidence in the disease state model (the hazard of onset of disease). In other cases, the interesting aspect is the state, corresponding to the prevalence in the disease state model (the probability of being in the state sick).

In order to discuss these aspects, we need to define the time, i.e., define a point, say time zero, from which the times are measured. Time is considered a positive, real-valued variable and thus has a continuous distribution. The time zero point needs to be known at the start of observation. As a counter example, a medical doctor might be tempted to define time zero as the time of diagnosis of the disease studied in order to evaluate how long the patients have symptoms before the diagnosis. However, this is unacceptable, as the time scale covers negative times. Individuals who have symptoms but

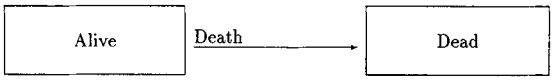

Figure 1. Lifetime state model.

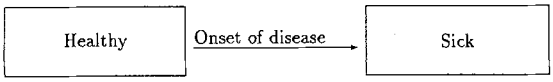

Figure 2. Disease state model.

do not develop the disease are implicitly ruled out from participating. To study the problem, a completely different set-up is needed, one where a group of individuals is followed and the response time is the time to diagnosis, measured in relation to their age, letting symptoms be described by a time-dependent explanatory variable. In most cases, the time zero will be the time of some event, i.e., a transition. When we measure time as age, the defining time point is birth, the time of transition from fetus to human being. For studying the occurrence of complication for a disease, the natural time scale is the duration of the disease, but for many diseases, this is impossible because the exact time of occurrence of disease is unknown. Instead, we use time since diagnosis of the disease, which is operational but not always scientifically satisfactory. For a drug trial, the natural time zero is start of treatment. Time zero does not need to be the time of start of observation; when it is not, the data are truncated (see Section 3.2). Table 1 gives a list of possible definitions of time zero. As an example of an entry into state, consider occupational mortality, where time zero is the time of start within the occupation. Another example is the female risk of heart disease, which is known to increase after menopause, not only due to age. In that case, one could define time zero as the time of menopause. Duration of pregnancy is traditionally measured since last menstrual bleeding because this time is known. However, the term is not intuitive, as conception takes place around day 10–14 within the pregnancy. As a time scale, it is acceptable because, when a woman is considered as pregnant, time zero is known, but it is necessary to be careful in the early phase; e.g., it makes no sense to consider mortality for women within the first week of pregnancy. Many medical studies are centered around a base line measurement on a group of people having a given disease at some point in time. From a scientific point of view, it can be natural to use age or duration of the disease as the time scale, with truncation, but in some cases, we might prefer to define time as time since the baseline measurement. A typical

Table 1
Possible choices of time scales

Time 0	Time scale
Birth	Age
Diagnosis of disease	Duration
Entry into state	Waiting time
Bleeding	Duration of pregnancy
Start of treatment	Length of treatment
Baseline measurement	Calendar time

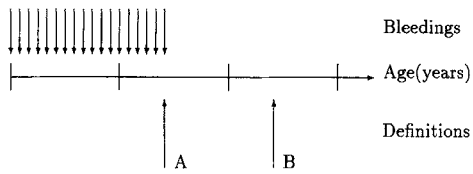

Figure 3. Menopausal model.

reason for doing so is lack of knowledge of methods to handle truncation. For most models, the definition of time zero is crucial, the only exception being constant hazard models.

In order to use survival data methods, the state must be known at all times until the end of observation. This is to be understood so that the current state must be known at all times, and this is certainly not trivial; e.g., menopause is defined as the time of the last menstrual bleeding, illustrated in Figure 3. Here, A marks the time of the last bleeding. But, at time A, we do not know that. It is not possible to say whether a bleeding is the last unless you have a whole lifespan. An operational definition is to wait for 1 year (time B) before we say that the bleeding was the last. At time B, we conclude that menopause happened at time A. The strict consequence is that any women who has had a bleeding within the last year does not know whether she has obtained menopause, even if she is only 20 years old. To become acceptable from a survival data viewpoint, menopause must be defined as the end of the bleeding-free year (time B), so that menopause is the first time she has had a year without bleedings. Using definition B, we can tell whether the woman has passed menopause at that time. It appears easy to add or subtract 1 year to go from one definition to the other, but this is not possible for all women, as the woman might die after a half year without bleeding, in which case it is forever unknown whether she should count as having obtained menopause (using definition A). Within the field of stochastic processes, the acceptable times are known as stopping times. Relaxing the requirement that the state is known at all times is possible, e.g., interval censoring, where the state is only known at some time points.

Survival data consider only a single or a few types of states and events, ruling out the possibility of continuous measurements of, e.g., quality of life. To study such a concept, it must be classified into a few well-defined groups. On the other hand, the time to the event is considered in great detail in survival analysis, sometimes more than relevant for the application. The biostatistical tradition for times for human beings is to count days and then analyze them by rank-based methods. For example, for studies of cancer, this does not make medical sense because cancers can be developing for up to 20 years before being diagnosed. Grouping cancers into years might be acceptable from a medical point of view but is unacceptable as the rank-based methods might have too many coinciding observations (ties). Similarly, in the economic example above, the hourly payment might be of higher importance to the individual than the time of obtaining the job, but survival analysis methods put the emphasis on the time to obtaining the job and can, at most, classify the salary into a few groups.

There are a number of books covering survival analysis, ranging from the comprehensive mathematical (Andersen et al., 1993) to the concise (Cox and Oakes, 1984) to the applied (Klein and Moeschberger, 1997). A book covering the more

advanced types of data, like multivariate and life history data, is in preparation by the present author.

3. Special Features of Survival Data

In broad terms, what makes survival data special is that the responses are times and thus are not measured in the same way as other variables. Other variables are measured almost instantaneously and independently of the response size. In survival data, larger observations take longer times to observe than smaller ones. Time is observed sequentially. This has several consequences. Most statisticians would mention the presence of censored data (Section 3.1) as the most important consequence. I think, however, that the concept of conditioning (Section 3.2) is much more fundamental. Other, less important considerations include the choice of parametric models (Section 3.3) and the possibility of having no or multiple events (Section 3.4). The possibility of time-dependent covariates (Section 4.2) is also a special feature of survival data.

3.1 *Censored Data*

Censored data means that the observations are only partially known. A common reason is that the person studied is alive when the study is evaluated, and thus all that is known about his lifetime is that it exceeds the age he had at the time of evaluation. The presence of censored data is a major technical problem. The occurrence of censoring must be unrelated to the future lifetime, i.e., the probability distribution of the residual lifetime for those censored must equal that of those who are not censored. If there are explanatory variables, the probability distribution of the residual lifetime of those censored must equal that of those who are not censored, having the same value of the explanatory variables. In almost all survival data cases, censoring is right censoring, i.e., observations are known to be larger than some given value. To combine expressions for events and right-censored observations, we introduce an indicator variable D, being one if the observation is an event and zero if it is a censoring. It might seem complicated to analyze such data, but this is not the case. The likelihood function includes terms corresponding to what we observe and thus, if an exact lifetime is observed, it contributes with its density and, if it is censored at some time t, it contributes with the probability that the lifetime exceeds t. Another reason for censoring is that the person is lost to follow-up, e.g., if a person emigrates, it is impossible to keep track of him. Even if it is possible to track the person, it can be irrelevant to include him if we study a risk factor related to the area that he moved from. It might be decided that observation ends at some prespecified time, e.g., if we are particularly interested in mortality before age 70, we can decide to censor the survivors at that age. If the aim is to study death from cardiovascular disease, the person is censored if he dies from cancer.

Censoring is not only a problem for survival data. Any measurement device has a range within which it functions and outside of which it only says that the result is outside of the range. For example, a thermometer can only measure the temperature within a given range.

3.2 *Conditioning*

This problem is most easily introduced by an example. We study the mortality data of the Statistical Yearbook (Danmarks Statistik, 1996) of Denmark with a few approximations, such as assuming constant density within each year and extrapolation above 100 years. According to the table, the median lifetime of a male is 75 years and 80 days, meaning that 50% die before that age and 50% after, based on the mortality experience in 1993–1994. Can a person celebrating his 75th birthday utilize this information, suggesting that he only has 80 days left, as evaluated by the median? No, we should take his present state as known, i.e., condition on him being alive at age 75. We can then evaluate the conditional probability that he will die within the coming 80 days (1.4%) and the conditional median given survival until age 75, which is 82 years 294 days. In contrast, if he had died at age 70, say, we would never ask the question of his lifetime because we already know the answer. What is relevant is the truncated distribution of the lifetime after age 75 years, i.e., the distribution given that the lifetime exceeds 75 years. This aspect has enormous consequences for the approach to survival data. One logical consequence is to discuss the distribution of the lifetime, say T, by means of the hazard function defined as the probability of death within a short interval, given that the person was alive at the beginning of the interval, i.e.,

$$\lambda(t) = \lim_{\Delta t \searrow 0} \frac{Pr(t < T < t + \Delta t \mid t < T)}{\Delta t}, \qquad (1)$$

which equals $-d \log S(t)/dt$, where $S(t)$ is the survivor function, i.e., $S(t) = \Pr(T > t)$, the probability of the lifetime being longer than t. The hazard function is well defined for continuous distributions. For survival data, the survivor function is more convenient than the ordinary distribution function $F(t) = 1 - S(t)$. The density is $f(t) = -dS(t)/dt$. The relations between the density and the hazard function are $\lambda(t) = f(t)/S(t)$ and $f(t) = \lambda(t) \exp\{-\Lambda(t)\}$, where $\Lambda(t) = \int_0^t \lambda(u)du$ is the integrated hazard function. Table 2 shows how the density, the survivor function, and the hazard are changed by truncation at time v, i.e., conditioning on $T > v$. This shows the advantage of the hazard, which, unlike the other quantities, is not changed due to the conditioning. It is already conditioned on survival until time t, and therefore it makes no difference to also condition on survival until time v ($v < t$). The hazard describes any aspect of the probability distribution. For example, a steeply increasing hazard corresponds to low variability and a decreasing hazard corresponds to high variability.

This conditioning corresponds to the concept of being at risk, i.e., before the event, the subject has to be in a state from where the relevant transition is possible. For example, you are

Table 2
The influence of truncation ($t > v$) on distributional quantities

Quantity	In full distribution	In truncated distribution given survival to time v
Survivor function	$S(t)$	$S(t)/S(v)$
Density	$f(t)$	$f(t)/S(v)$
Hazard function	$\lambda(t)$	$\lambda(t)$

only at risk of death if you are presently alive. You cannot contract a disease you already have. Some nonparametric estimation methods highlight the conditioning principle as successive conditioning, where at each time of event, the method conditions on the persons being alive immediately before the time point.

From a mathematical point of view, this is naturally analyzed as a random process developing over time, i.e., a stochastic process, with one realization of the process for each individual. As the response is a single time, the stochastic process to choose is a counting process, being zero before the event and one at and after the event. This seems a rather complicated way of describing a random variable, but it is necessary in order to accommodate the special features of survival data.

The discussion on truncation has concerned what quantities we would like to estimate. But truncation is also relevant for the data. Very often, the time of start of observation equals time zero, and thus truncation is not a problem, but it might be more relevant to let the time be described by the age and then start observation at whatever age the person is at the start of the study, say t_0. This is truncation at time t_0 and is called late entry.

3.3 *Choice of Parametric/Nonparametric Models*

The standard distributions we usually apply to other types of data are, in most cases, not relevant to survival data. This is illustrated by trying to fit such distributions to a lifetime distribution. As described above, the survival time is necessarily positive. For many other types of positive data, we apply the normal distribution, even though we know that it is in conflict with the data being positive. One such example is the height of human beings. With the relevant values of the mean and variances, the probability that the normal distribution suggests a negative height is so small that we need not care. It may or may not be the same for survival data, but the possibility of negative values seems particularly awkward in this case.

3.3.1 *Illustration using population data.* This section is an illustration of how an actual analysis with standard methods would go. Figure 4 shows the density for the data on Danish population mortality. We show the density rather than the hazard in order to make a standard analysis. The mean is 72.5 years and the standard deviation is 15.5 years for males, and the corresponding quantities are 77.8 and 14.9 years for females. If we apply the normal distribution, we evaluate that the probability of a negative lifetime is 0.0002% for males. Thus, negative lifetimes do not create a major problem. But we can assure that the variable is positive valued by the standard approach of applying the normal distribution after a logarithmic transformation or applying some other well-known continuous distribution on the positive numbers, like the gamma or the inverse Gaussian. The latter is the distribution of a waiting time in a Brownian motion and is therefore theoretically interesting as a distribution for a time variable. However, all three suggestions lead to very bad fits because these distributions have positive skewness, where human lifetimes have negative skewness. Figure 5 illustrates this. The distribution from the Statistical Yearbook (Danmarks Statistik, 1996) is shown for males along with normal, log-normal, gamma, and inverse Gaussian distributions fitted by the method of maximum likelihood on a theoretical data set generated from the distribution of the table. Infant mortality influences all estimates and is one reason for the bad fit both during infancy and later in life. The models suggest very small mortality during infancy and childhood and that there is a very high proportion of long-term survivors, ranging from 0.8% living past 110 years in the normal distribution to 18% for the inverse Gaussian, in contrast to the true value, which is in the order of 0.001%.

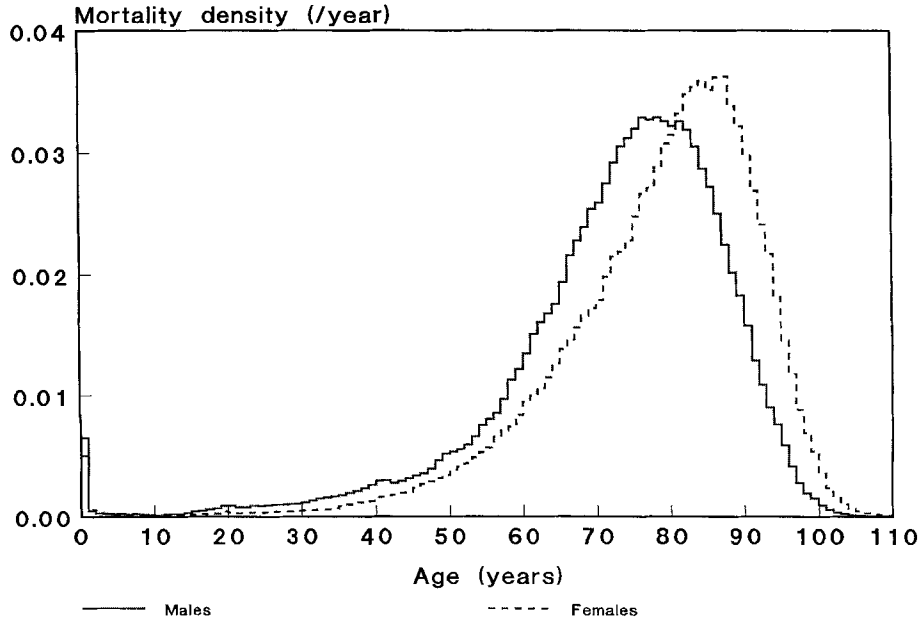

Figure 4. Density of lifetimes in the Danish population.

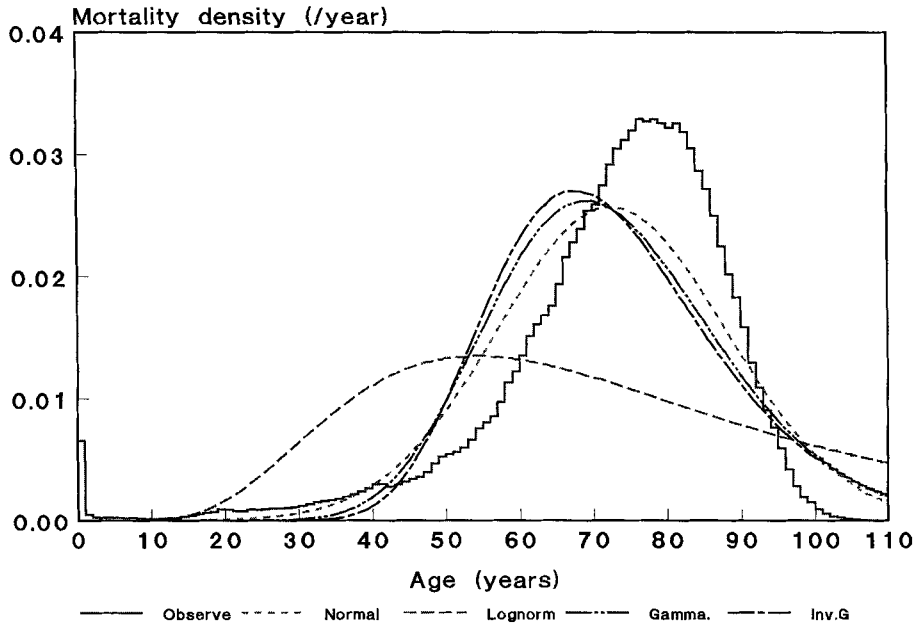

Figure 5. Fitted standard densities to lifetimes in the Danish male population.

3.3.2 *Parametric survival distributions.* We need distributions that are positive valued but show negative skewness. Two such distributions are the Weibull and the Gompertz distributions. The Weibull has hazard $\lambda(t) = \mu\gamma t^{\gamma-1}$, $\mu > 0$ and $\gamma > 0$, with decreasing hazard for $\gamma < 1$ and increasing hazard for $\gamma > 1$. The mean and variance can attain any positive values. The Gompertz has hazard $\lambda(t) = \mu\varphi^t$, $\mu > 0$ and $\varphi \geq 1$, with increasing hazard for $\varphi > 1$. The mean can attain any positive value, and the standard deviation is less than or equal to the mean. Inserting $\varphi < 1$ leads to a decreasing hazard with a positive probability of an infinite lifetime. Both families include constant hazards (exponential distributions) as $\gamma = 1$, respectively $\varphi = 1$. Both of them have negative skewness when the variability is low. For the Weibull, this happens when γ is above approximately 3.6, corresponding to a coefficient of variation below 0.308. For the Gompertz distribution, a similar simple parameter relation cannot be specified, but the skewness is negative when the coefficient of variation is below 0.495. These distributions easily adapt to truncation. In fact, the residual lifetime after truncation of a Gompertz distribution is also a Gompertz distribution with the same value of φ. The term residual means that the time is restarted at the time of truncation, i.e., the truncation time is subtracted from the original time. Furthermore, in order to fit the child mortality, the limit of the density at $t = 0$ should be positive. The log-normal, the inverse Gaussian, the gamma (for most parameter values), and the Weibull ($\gamma > 1$) distributions do not satisfy this requirement as the limit of the density is 0 at 0. The Gompertz distribution has positive density at 0 and gives a fine fit to the population data except for the first couple of years.

3.3.3 *Nonparametric methods.* Nonparametric methods have become very popular within survival analysis for several reasons. One reason is that data often have some features that are not easily obtained by parametric models. For example, human lifetimes show a decreasing hazard the first 5 years of life. This feature is relevant for some considerations and irrelevant for other, but by assuming a nonparametric distribution, one avoids having to think about this problem. Similarly, there is an increased hazard the first few years after the legal driving age (18 years in the country studied). For many other types of data, there are also some areas where we know that the fit of a parametric model is bad, but in survival data, these are tied to particular ages, and therefore we find it more inconvenient. A more important reason is related to the conditioning discussed earlier. Consider again the 75-year-old male. How should we evaluate his chance of surviving to age 76? It is very natural to say that this should be based on the survival experience of previous males of age 75. In a parametric model, this is instead based on all the information in the data set, including, e.g., males who died at age 20. In that sense, the nonparametric methods can be considered the logical consequence of the conditioning principle applied to the data. Above, the conditioning principle was used to illustrate which quantities were relevant, and it leads to parameterizing the lifetime distribution by means of the hazard function. Now, it has further consequences, as it tells how to make inference on these parameters. Nonparametric models have several shortcomings. If a concrete data set on males of age 75 includes only a single death before the age of 76, say at the age of 75 years and 10 days, then we estimate that it is absolutely impossible to die at the age of 75 years and 11 days, which is in conflict with common sense. A reasonable compromise between the simple parametric models and the completely nonparametric models is the piecewise constant hazard model, where the hazard is assumed constant in specified intervals.

In industrial applications and demography, it is common to evaluate the mean lifetime, as done above. However, within

biostatistics, this is considered unacceptable. One reason is that censoring generally makes it difficult to estimate the right tail, and this tail can have a marked influence on the mean. For the population data, we have demonstrated that the data are left-skewed, implying that assumptions made about the right tail have smaller influence. This argument does not carry over to other times than lifetimes. A second reason is that evaluation of moments might make people think of the response as normally distributed, which can be quite misleading. Finally, as described below, for some types of events, there is a proportion never experiencing the event, which invalidates the use of moments.

3.4 *Multiple Events*

To apply ordinary methods for survival data, we need to define some random variables. For true lifetimes, this is easy because the lifetime is well defined, even when it is not observed due to censoring. All people eventually die. However, for many other events, we cannot be sure that the event will ever happen. A person may or may not develop a specified disease even with long observation time. Thus, in order to define a random variable corresponding to that disease, we must allow a value of infinity. This is not appropriate for the standard models. For other diseases, an individual might get multiple attacks of the same disease, making it impossible to fit into a standard set-up of univariate random variables. The hazard-based set-up for survival data allows the possibility of there being zero or many events for each individual.

4. Regression Models

In many cases, explanatory factors or covariates are available. These might be factors that are of intrinsic importance for the application, like treatment in a drug trial, or factors that are known or suspected to influence the hazard of event but that are not interesting *per se*. In fact, when there are covariates, we are typically more interested in the effect of those than in how the hazard changes over time. Let the p-vector of covariates for a person be $z = (z_1, \ldots, z_p)$. By far, the most common model is the proportional hazards model, specifying that the hazard for a person with covariate z is

$$\lambda_0(t) \exp(\beta' z). \tag{2}$$

Thus, the hazard is a product of a term depending on time and a term depending on the covariates. The parameter β describes the importance of the covariates. Cox (1972) suggested an estimation procedure that removed the effect of the term $\lambda_0(t)$, letting it be completely unspecified. Thus, the analysis concentrates on the effect of the covariates.

Rather than having the hazard function be described as a function of the explanatory variables, it is possible to let the explanatory variables act directly on the time via a scale factor. This is the accelerated failure time model. Thus, there is a basic survivor function $S_0(t)$, the survivor function for a person with covariate value 0, and a survivor function for a person with covariates z, i.e., $S_0\{t/\exp(\eta' z)\}$. The parameter η describes the importance of the covariates. If $S_0(t)$ corresponds to a Weibull distribution of shape γ, this gives the same regression model as the proportional hazards model but with a different parameterization. The parameter relation is

$$\eta = -\beta/\gamma. \tag{3}$$

There are no other models where the proportional hazards and the accelerated failure time models are identical. The standard way to describe an accelerated failure time model is as

$$\log T_i = \eta' z_i + \epsilon_i, \tag{4}$$

where the ϵ_i are independent and identically distributed.

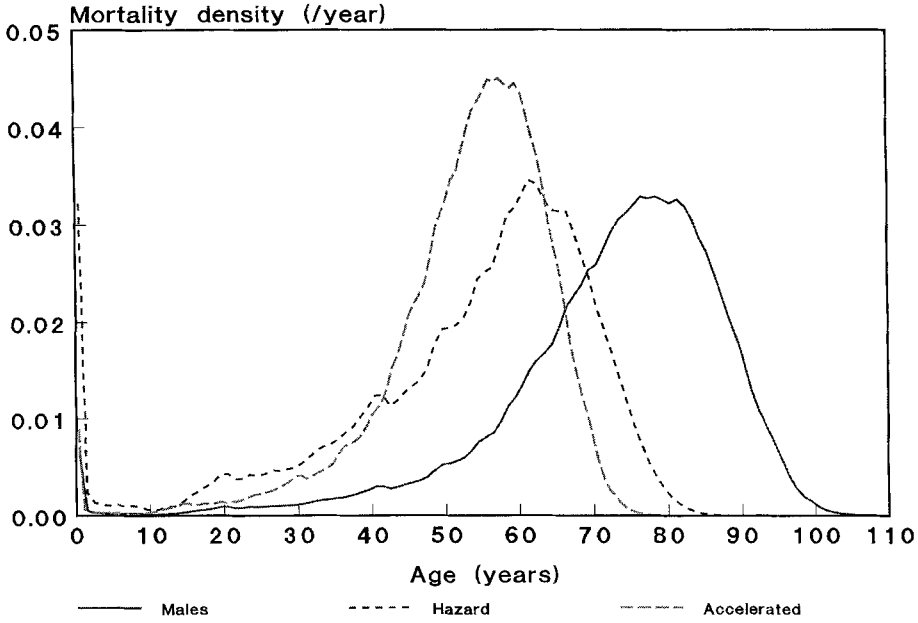

Figure 6. The density for the mortality of Danish males before and after modification by proportional hazards (factor 5) and accelerated failure time (factor 0.732).

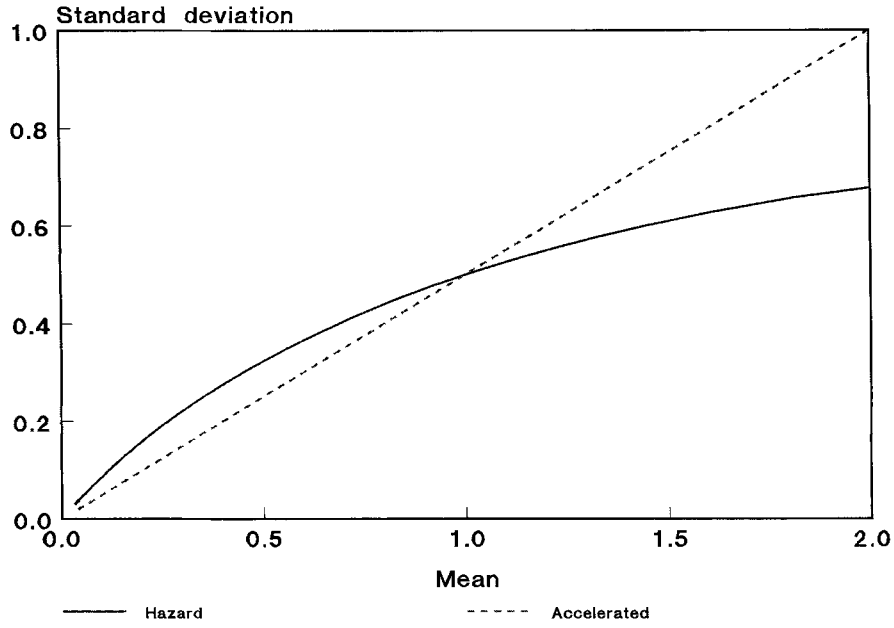

Figure 7. The relation between the mean and standard deviation in the proportional hazards and accelerated failure time model, based on a Gompertz distribution with mean 1 and standard deviation 1/2.

Thus, this describes a linear model for the logarithm to the times. The distribution of the error ϵ_i can be of Weibull form or of some other form but is typically parameterized.

4.1 *Comparison of Accelerated Failure Time Models and Proportional Hazards Models*

The question is, of course, whether the accelerated failure time model makes more or less sense than the hazard regression model. Accelerated failure times seem more sensible for wear-related processes, i.e., if a constant load differs between individuals. In some cases, this can be interpreted as a double time scale. The lifetime of a car might be measured in years but could maybe be better described by the mileage. Then the load could be the distance driven per year. The distance is, however, typically not known and therefore a regression model is introduced in order to describe the distance as a function of the covariates. A hazard-based model on the other hand makes it easier to accommodate time-dependent covariates. It is also better for describing changes in risk corresponding to fixed points in the time scale, e.g., the change in risk for boys when they attain the age to receive a driver's license. A main advantage of the proportional hazards model is that an estimate can be found allowing for an arbitrary hazard function, where the accelerated failure time model typically is based on a parametric model.

On the graphical side, this can be understood on the integrated hazards scale. Suppose there are two groups. The proportional hazards model then says that $\Lambda_2(t) = c\Lambda_1(t)$ for some c, whereas the accelerated failure time model says that $\Lambda_2(t) = \Lambda_1(t/c)$ for some c.

This can be illustrated by considering the effect on a chosen hazard. Take the distribution for the Danish male population depicted in Figure 4. If the hazard is multiplied by 5, the mean is reduced from 72.5 to 53.1 years. Multiplying on the time axis by 0.732, corresponding to an accelerated failure time model, leads to the same mean lifetime. The effect on the actual distribution is seen in Figure 6, which illustrates the density for the original distribution and the two modified distributions. The accelerated failure time approach leads to a decreased mean with a constant relative variability and thus a decreased absolute variability, clearly seen by the higher and more narrow peak density. The proportional hazards approach similarly leads to a decreased mean but, neglecting low ages, it rather looks like a translation, and the variability appears to be unchanged. For the infant mortality, there is a clear change for the proportional hazards model but no big difference for the accelerated failure time approach.

This can be supported theoretically by the Gompertz distribution. The proportional hazards model modifies a Gompertz distribution into another Gompertz distribution with a different value of μ but leaving φ untouched. The accelerated failure time model also leads to another Gompertz distribution, but both parameters are changed. This is illustrated in Figure 7, where the baseline is a Gompertz distribution of mean 1 and standard deviation 1/2. The figure shows the means and standard deviations obtainable in the two models. In the accelerated failure time model, the standard deviation is proportional to the mean, as the modification is a scale factor on time. For the proportional hazards model, the standard deviation changes less than the mean value, similar to the results found for the population data.

4.2 *Time-Dependent Covariates*

When introducing the proportional hazards model, Cox noted that the covariates could depend on time, the only requirement being that $z(t)$, the value of the covariate at time t, be known at time t. This is not always satisfied, as will be

discussed below. Time-dependent covariates are very well suited to the successive conditioning and are mathematically simple to handle but conceptually difficult to understand fully because they offer an extremely powerful modelling tool. The likelihood is the same except that the value of z is exchanged by the value of z evaluated at the death time. One has to apply time-dependent covariates with care in order to get useful results. A key point is whether the covariates can be predicted. For example, if covariates monitoring the heart function, such as pulse and blood pressure, are included, they will often change immediately before death and then explain the death in the sense they have high predictive power. If we want to study the effect of a cardiovascular drug treatment, we might not directly see the effect if the drug acts by stabilizing pulse and blood pressure. This also makes it impossible to predict death well in advance because we do not control how pulse and blood pressure change. To handle these problems, pulse and blood pressure should be included as responses in the model, i.e., a more detailed model for the life history must be made. A more extreme example is the menopausal model of Figure 3. Defining menopause at time B, the time when the woman has experienced a bleeding-free year, which makes sense from a survival analysis viewpoint, we can make a perfect model by including a time-dependent covariate, measuring time since latest menstruation. As long as this covariate is below 365 days, the incidence of menopause is 0, and when it gets to 365 days, there is 100% probability that menopause happens. Even though there are many other factors that influence time of menopause, this time-dependent covariate takes the honor of describing it. All we can say when the covariate is z is that, if the woman does not have a bleeding within the next $365 - z$ days, she will obtain menopause at the end and that, if she bleeds, menopause will be more than a year from now. This is exactly what the definition of menopause says and is therefore not interesting. To make a nontrivial conclusion, we need to exclude such covariates. A similar example is that for studying births. The best predictor is a time-dependent variable describing pregnancy, but this is only good for short-term predictions of births, i.e., up to 9 months. To make a long-term prediction or to examine the effect of age, race, partner, and number of previous births, pregnancy must be excluded as a possible covariate. Another way to say this is that choosing a model is not only a matter of obtaining a good fit but that the predictability of the time-dependent covariates must also be taken into account. The covariate function must be left continuous to exclude response-dependent functions like $z(t) = 1\{T \leq t, D = 1\}$, which tells nothing but has a perfect fit.

It is possible that the covariate trajectory is known from the beginning, or it is possible that it changes in a more or less unpredictable way. Some key types of time-dependent covariates are given in Table 3. Some possible applications can illustrate the points. In a study of persons with a given disease, it is relevant to ask whether it is the age, say a, or the duration of the disease, say u, that determines mortality. Both can be included when time-dependent covariates are allowed for. We can pick age as the time scale in a model with hazard $\lambda_0(a)\exp(\beta u)$. As duration equals time since age at diagnosis (a_0), this can be reformulated as

$$\begin{aligned}\lambda_0(a)\exp\{\beta(a-a_0)\} &= \lambda_0(a)\exp(\beta a)\exp(-\beta a_0)\\ &= \tilde{\lambda}_0(a)\exp(\tilde{\beta}a_0).\end{aligned}$$

Here the arbitrary function $\lambda_0(a)$ can absorb any function of a, simplifying the model to a fixed covariate model, with age at diagnosis as covariate, and $\tilde{\beta} = -\beta$ as regression coefficient and with truncated data. Alternatively, we can pick duration as the time scale to make a model with hazard $\mu_0(u)\exp(\kappa a)$. This model can similarly be transformed to a fixed covariate model with hazard $\tilde{\mu}_0(u)\exp(\kappa a_0)$. In this case, the regression coefficient is unchanged. There is no truncation if the patients are included from the time of diagnosis. This transfer of the time-dependent terms into $\lambda_0(t)$ works for any common function (implying that such terms cannot be fitted) and for linear functions with common slope. Whether we should pick age or duration as the time scale is a matter of fit. We should choose the variable with the largest effect or the most nonlinear effect. General known covariates can be used to model nonproportional hazards. For example, a nonproportional treatment effect (a treatment by time interaction) can be described by a fixed variable z_1 being the indicator function of treatment and a time-dependent term $z_2(t) = tz_1$. Thus, the hazard in the control group is $\lambda_0(t)$ and in the treatment group $\lambda_0(t)\exp(\beta_1 + \beta_2 t)$. This gives the same relative risk function as a Gompertz model with different values of φ in the two groups. In this model, β_1 describes the initial treatment effect and β_2 the change in treatment effect over time. This is particularly relevant for testing proportional hazards, which is the submodel $\beta_2 = 0$, but less relevant for modelling nonproportional hazards, as the consequence of the assumption is that the treatment effect is reversed if β_1 and β_2 have opposite signs and that the effect either increases to infinity or decreases to minus infinity with time. To avoid such problems, one can pick covariates that stay finite, e.g., piecewise constant variables. The covariate can further be an indicator function of presence in a given state. This is a general approach and can be used for modelling of the hazard function by, e.g., the presence of complications. What makes this type of variable special is that there are only a few possible states, and the transitions between the states can be included in the model. A classical example is the survival among the participants in the Stanford Heart transplant program (Crowley and Hu, 1977), where the hazard of death differs before and after heart transplant. The final type to mention is that of measured variables, which means that the covariate trajectory is external to the model and thus impossible to predict. These may be measured in a

Table 3
Some types of time-dependent covariates

Type	Application
Known linear, common slope	Multiple time scales
Known	Check of proportionality assumption Modelling nonproportionality
State-dependent	Multistate models
Measured	Dependence examination

regular pattern or at visits made with irregular intervals. If the intervals are long and there is a clear development over time, which, e.g., would be the case for CD4 counts in AIDS and for albuminuria measurements in diabetes, it seems attractive to use interpolation to get the most precise evaluation of the variable at any given time. This, however, invalidates the principle that the value of $z(t)$ is known at time t. It should, in fact, be clear that this is wrong because it is not possible to use interpolation for those that die, and therefore the covariates are treated differently for the deaths and the survivors. If the simple approach of using the latest measurement is not satisfactory, it is possible to make an updating formula based on previous values. If there are many measurements, we can use extrapolated values based on a fit of earlier measurements, but in other cases, we can modify the latest measurement by some function of the time since the measurement. For the albuminuria example above, the concentration is known to increase by about 15% per year as a mean over patients, and we can then use this number for extending the value.

For predictions, it is necessary to know the future course of the covariates, either precisely or as a distribution of the covariates. This is not a problem for known covariates, where the course is fixed from the beginning. In the case of state-dependent covariates, the future development can be modeled as a random quantity. In the case of measured covariates, we have to make an assumed development for the covariate process or set up a stochastic process for its future, and this might be difficult or impossible.

Time-dependent covariates are easily introduced in parametric proportional hazards models. For the likelihood, we need the hazard at the time of event and the integrated hazard over the observation period, which is easily found. Time-dependent covariates can also be introduced in accelerated failure time models, but this is more complicated than for proportional hazards models. We should define an operational time $v(t)$, which is the time for a given trajectory $z(t)$, measured in standard person time units (i.e., corresponding to $z(t) \equiv 0$). The operational time should move with a speed of $\exp(-\eta' z(t))$ compared to real time and thus is defined by $v(t) = \int_0^t \exp(-\eta' z(u))du$. This implies that the survivor function for a given person should have the form

$$S(t \mid z(u), 0 \le u \le t) = S_0\left(\int_0^t \exp\left(-\eta' z(u)\right) du\right), \quad (5)$$

where $S_0(t)$ is the survivor function for a standard person. This definition is only valid if the covariate trajectory is known from the beginning because it requires the existence of $z(t)$ for the whole interval, including for persons dying before t. To generalize the expression, we make it hazard based, i.e.,

$$\lambda(t \mid z(u), 0 \le u \le t) = \lambda_0\left(\int_0^t \exp\left(-\eta' z(u)\right) du\right) \exp\left(-\eta' z(t)\right), \quad (6)$$

where $\lambda_0(t)$ is the hazard for a standard person. This model is the same as (5) but only requires $z(t)$ be known when the person is alive. This is the same model as the proportional hazards time-dependent covariate model only when $\lambda_0(t)$ is constant. In the special case where $z(t) = ct$, with c an individual known factor ($c > 0$), $\eta < 0$, and $\lambda_0(t)$ constant, this model becomes a Gompertz model with $\varphi = \exp(-\eta c)$.

4.3 *Neglected Covariates*

In some cases, we cannot include all relevant covariates due to lack of knowledge. From the linear normal model, it is known that, when the unknown covariates are independent of the known covariates and follow a normal distribution, the distribution of the response is still normal, but the variance is increased. The problem with standard survival models is that typically neglected covariates lead to distributions outside the family considered. For example, considering neglected covariates in a constant hazard model leads to a distribution with decreasing hazard. The accelerated failure time model is formulated as

$$\log T = \eta' z + \kappa' w + \epsilon, \quad (7)$$

where z are the observed covariates and w the neglected covariates. As w is unknown, we have to assume it random and integrate it out. This means that the distribution of ϵ is changed into one with larger variability and is typically outside the parametric family considered. The positive point is, however, that there is no change in the regression part of the model, implying that the coefficients η are unchanged. Unfortunately, a similar result is not true for the proportional hazards case. To be specific, we generalize equation (2) to

$$\lambda_0(t) \exp\left(\beta' z + \rho' w\right). \quad (8)$$

It is sufficient to describe the distribution of the univariate function $Y = \exp(\rho' w)$, the so-called frailty. The nicest case is when Y follows a positive stable distribution of index α, $\alpha \in (0, 1]$. Then the hazards, after integrating out the neglected covariates, are still proportional as a function of the observed covariates, but the regression coefficient β is reduced to $\alpha\beta$, i.e., closer to zero (Hougaard, 1986). In the Weibull case, the two models coincide and the distribution is still Weibull, but the shape parameter is reduced. So in this case, the distribution stays within the family when the neglected covariates are integrated out. The key equation is (3), which is consistent with the η parameter being unchanged but β not being, showing that the accelerated failure time parameterization is a better one than the proportional hazards parameterization. For all frailty distributions other than the positive stable, the relative risk is no longer a constant function (i.e., the hazards are not proportional) and the relative risk between two persons is closer to one than in the conditional distribution given the unobserved covariates. For example, when Y follows a gamma distribution, the relative risk at time zero equals the value $\exp(\beta'(z_2 - z_1))$, expected from equation (2), but the relative risk converges to one as time goes to ∞. When Y follows an inverse Gaussian distribution, the relative risk instead converges to $\exp(\beta'(z_2 - z_1)/2)$ as time goes to ∞. Without observed covariates, it is also possible to have neglected covariates in model (8) and stay within the Gompertz family (Hougaard, 1986).

In summary, the accelerated failure parameter η is robust toward neglected covariates, whereas the proportional hazards parameter β is not. An application to the incidence of diabetic nephropathy, where this effect is clear, is presented by Hougaard, Myglegaard, and Borch-Johnsen (1994). It is a major drawback of the proportional hazards model that the model and the value of the relative risk are not robust toward neglected covariates.

Résumé

Les données de survie se distinguent par une discipline statistique particulière. Cet article essaie de décrire ce que sont les données de survie et ce qui les rend aussi particulières. Les données de survie concernent des délais jusqu'à l'apparition de certains événements. Un point crucial est l'observation du temps en succession, qui d'un côté mène à la non-observation de certains délais, tout ce qui est connu est qu'ils sont supérieurs à certains délais donnés (la censure), et de l'autre côté implique que les prédictions sur le cours futur doivent être conditionnées sur l'état actuel (la troncature). Dans le cas le plus simple, cette condition implique que l'individu soit vivant. Le conditionnement successif fait en sorte que la fonction de risque est le concept le plus pertinent, car elle décrit la probabilité qu'un événement ait lieu dans un petit intervalle de temps, sachant que l'individu est en vie aujourd'hui (ou plus généralement apte à subir l'événement). Les distributions standards disponibles (normale, log-normale, gamma, Gaussienne inverse etc.) peuvent prendre en compte la censure et la troncature, mais ceci est difficile à manier. De plus, elles s'ajustent mal parce qu'elles sont soit symétriques, soit asymétriques vers la droite, alors que les distributions des délais de survie peuvent être facilement des variables positives asymétriques vers la gauche. Quelques distributions remplissant ces conditions sont disponibles, cependant des méthodes non-paramétriques sont souvent préférables puisqu'elles prennent mieux en compte la troncature et la censure de façon conceptuelle, et elles donnent une meilleure adéquation. Enfin, on compare les modèles de régression à risques proportionnels aux modèles accélérés des délais de survie.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer Verlag.

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. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association* **72,** 27–35.

Danmarks Statistik. (1996). *Statistical Yearbook.* Copenhagen: Danmarks Statistik.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73,** 387–396. (Correction **75,** 395.)

Hougaard, P., Myglegaard, P., and Borch-Johnsen, K. (1994). Heterogeneity models of disease susceptibility, with application to diabetic nephropathy. *Biometrics* **50,** 1178–1188.

Klein, J. P. and Moeschberger, M. (1997). *Survival Analysis.* New York: Springer Verlag.

Received October 1997. *Revised January* 1998.
Accepted January 1998.