Survival analysis: Methods for analyzing data with censored observations

TITLE: Survival analysis: methods for analyzing data with censored observations

AUTHOR: Tomasz Burzykowski

AFFILIATIONS: Data Science Institute, Hasselt University, Agoralaan D, 3590 Diepenbeek, Belgium; Department of Biostatistics and Medical Informatics, Medical University of Bialystok, Szpitalna 37, 15-295 Bialystok, Poland

EMAIL: tomasz.burzykowski@uhasselt.be

ADDRESS FOR CORRESPONDENCE: Data Science Institute, Hasselt University, Agoralaan D, 3590 Diepenbeek, Belgium

Abstract

This article provides a review of basic statistical methods for the analysis of data that include censored observations.

Keywords: censoring, Kaplan-Meier estimator, logrank test, accelerated failure-time model, proportional-hazards model

Introduction

Censoring occurs when we do not observe exactly the value that we are interested in, but we only learn about some bounds surrounding it. In particular, an observation is *left-censored* when it is larger than the true value. A *right-censored* observation is smaller than the true value. *Interval censoring* occurs when we learn that the true value lies within the interval limited by two observed values.

Censoring is most often encountered in the case of observing a *time to event* (TTE), i.e., the time that elapses between a well-defined starting moment until a particular event of interest. For instance, in a retrospective cohort study organized in Iowa [1], dental records of 200 children younger than 6 years of age were used to investigate the age (in months) until the first dental caries. The records were followed for a minimum of 36 months after the first dental visit. In the study, left-censored observations were obtained for children with dental caries diagnosed at the first visit, and indicated that the age at the first caries was shorter than the age at the child's first dental visit. Right-censored observations were obtained for children, for whom no dental caries was diagnosed or treated at any time throughout the study; they indicated that the age at the first caries was larger than the age at the last recorded visit. Interval-censored observations were obtained for children for whom dental caries was diagnosed during the study period; they indicated that the age at the first caries was equal to a value within the interval defined by the age at the last visit at which no caries was diagnosed and the age at the first visit at which caries was diagnosed. Figure 1 illustrates the different types of censoring for a hypothetical set of eight children entering the study at different ages with three 6-month visits recorded for the study purposes.

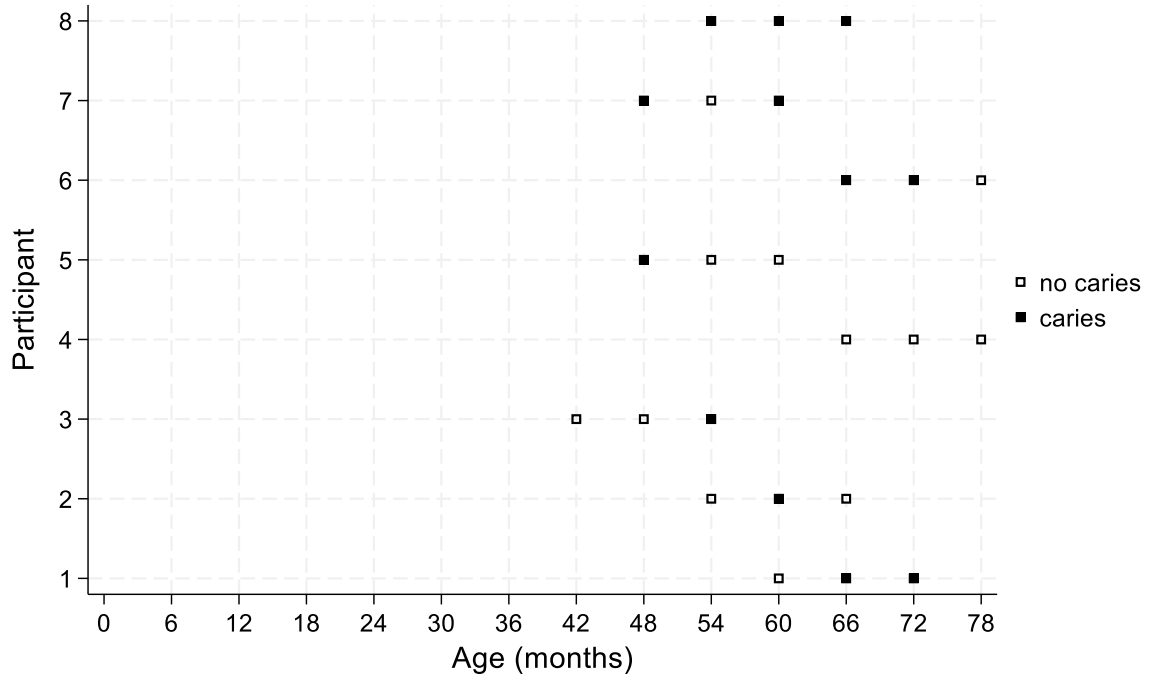

Figure 1. A hypothetical set of eight children entering the retrospective cohort study at different ages with three 6-month visits recorded for the study purposes. Visits with diagnosed caries are indicated by the filled-in square. For participants 1, 2, and 3, interval-censored observations are obtained that indicate that the time to the first caries lies in the intervals (60, 66) (54, 60), and (48, 54), respectively. For participant 4, a right-censored observation equal to 78 months is recorded. For participants 5, 6, 7, and 8, left-censored observations are obtained, equal to 48, 66, 48, and 54 months, respectively.

Note that censoring may apply to any measurement or observation, not only a TTE. For instance, left- and right-censoring applies to diagnostic assays with, respectively, lower and upper limits of detection. In the remainder of this article, however, we will focus on the case of observing a TTE. Moreover, we will consider analysis of data that include exact (uncensored) or right-censored observations of a TTE. This is because the presence of left-censored observations requires only a minor modification of methods of analysis of data with right-censored observations. On the other hand, in practice, interval-censored observations are very often (though incorrectly) transformed into uncensored ones by assuming that the event took place at the time equal to the observed upper limit of the interval. For instance, interval-censored observation of time to first caries is replaced by the time of the first visit, at which caries is diagnosed.

The presence of censored observations complicates the statistical analysis. This is because, in such a case, the use of the classical statistics (such as, e.g., the sample mean) or statistical models (such as, e.g., linear regression) will result in biased results. For instance, if some of the observations in a dataset are right-censored, the sample mean will underestimate the true mean of the TTE.

Analysis of data that include censored observations requires the use of methods that take into account the censoring. Collectively, in medicine, these methods are referred to as survival analysis. They can be parametric or non-parametric.

In *survival analysis*, we are interested in making statements about the true (unknown) distribution of the TTE by using the observed (and possibly censored) data. We may want to get an idea about, for instance, a particular characteristic of the distribution such as the mean or the median. Or we may be interested in getting an idea about the entire distribution by estimating its cumulative distribution function or its complement, the survival function.

Toward this aim, we may use parametric or non-parametric methods. *Parametric methods* estimate the characteristics of the TTE distribution by making concrete assumptions about the form of the distribution. For instance, we may assume that the TTE has an exponential or log-normal distribution. *Non-parametric methods* avoid making such assumptions. In general, parametric methods provide more precise estimates than the non-parametric methods. On the other hand, if the assumption about the TTE distribution is incorrect, the estimates obtained by applying a parametric method may be biased.

To illustrate various methods discussed in the article, we will use data from two motion-sickness studies [2]. In both experiments, a motion generator was used to study effects of various motion parameters on the risk of motion sickness at sea. Participants in the studies were observed with respect to the time of the first occurrence of frank emesis. Each of the experiments limited the exposure time to two hours, unless the participant requested to stop the experiment or experienced the event (emesis). Thus, for participants who did not experience the event for the entire period of two hours, or who prematurely quit the experiment, right-censored observation of the TTE was recorded. In one study, “soft motion” (with the frequency of 0.167 Hz and the acceleration of 0.111 G) was simulated. In the other one, “hard motion” (with the frequency of 0.333 Hz and the acceleration of 0.2222 G) was simulated.

The recorded times (in minutes) for the "soft motion" experiment, with 21 participants, are as follows (right-censored times are indicated by the asterisk): 30, 50, 50*, 51, 66*, 82, 92, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*. Note that this study yielded five uncensored observations and 16 right-censored ones (two for participants who quit the experiment prematurely at 50 and 66 minutes and 14 for the participants who endured the entire 120 minutes without an event).

The recorded times (in minutes) for the "hard motion" experiment, with 28 participants, are as follows (right-censored times are indicated by the asterisk): 5, 6*, 11, 11, 13, 24, 63, 65, 69, 69, 79, 82, 82, 102, 115, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*, 120*. This study yielded 14 uncensored observations and 14 right-censored ones (one for the participant who quit the experiment at 6 minutes and 13 for the participants who endured the entire 120 minutes without an event).

Parametric methods

We will discuss the use of a parametric approach by using the example of the *exponential distribution*. The distribution is governed by a single parameter, which we will denote by λ. If the TTE is generated by an exponential distribution, then the mean value of the time is equal to $1/\lambda$, the median value is equal to $\ln(2)/\lambda$, and variance is equal to $1/\lambda^2$. Thus, all characteristics of the TTE distribution depend on the single parameter. Consequently, if we can estimate λ, then we can subsequently derive from it estimates of the mean, variance, etc.

In case of an exponential distribution, the estimation of λ based on data that includes right-censored observations is very simple. In particular, we estimate λ by taking the ratio of the number of events to the sum of (all) observed values of the TTE. For instance, in the case of the "soft motion" study, there were five events, while the sum of all observed times was equal to

$$30 + 50 + 50 + 51 + 66 + 82 + 92 + 14 \cdot 120 = 2101.$$

Thus, the estimated value of $\lambda = 5/2101 = 0.0024$. Consequently, the mean TTE can be estimated to be equal to $1/\lambda = 2101/5 = 420.2$ minutes, i.e., for a person subjected to the "soft motion" we would expect a TTE of about 420 minutes. On the other hand, the estimated median TTE is equal to $\ln(2)/\lambda = 0.693/(5/2101)=291.2$ minutes, i.e., we would expect that half of persons subjected to the "soft motion" would experience the event before 291 minutes. Note that both of these values are much larger than the maximum time (120 minutes) observed in the study. Thus, these estimates are extrapolations beyond the range of observed the data, and their validity very much depends on the assumption of the exponential distribution.

For the "hard motion" study, the estimated value of $\lambda = 14/2356 = 0.0059$. Thus, it is higher than for the "soft motion" study. As a consequence, the estimated values of the mean ($1/\lambda = 2356/14 = 168.3$ minutes) and median ($\ln(2)/\lambda = 0.693/(14/2356)=116.6$ minutes) are smaller than for the "soft motion" study. This suggests that the increase of the frequency and the acceleration shortened, on average, the TTE.

Instead of estimating a particular characteristic, we may be interested in characterizing the entire distribution. Toward this aim, in survival analysis, we often focus on the *survival function* (sometimes also called *survivor function*). Traditionally, the function is denoted by $S(t)$. For a particular value of the argument t, the function provides the probability that the TTE will be larger than or equal than t. Or, in other words, the probability that the event of interest will not take place by time t. In case of death, this is the probability of surviving at least time t; hence the name of the function.

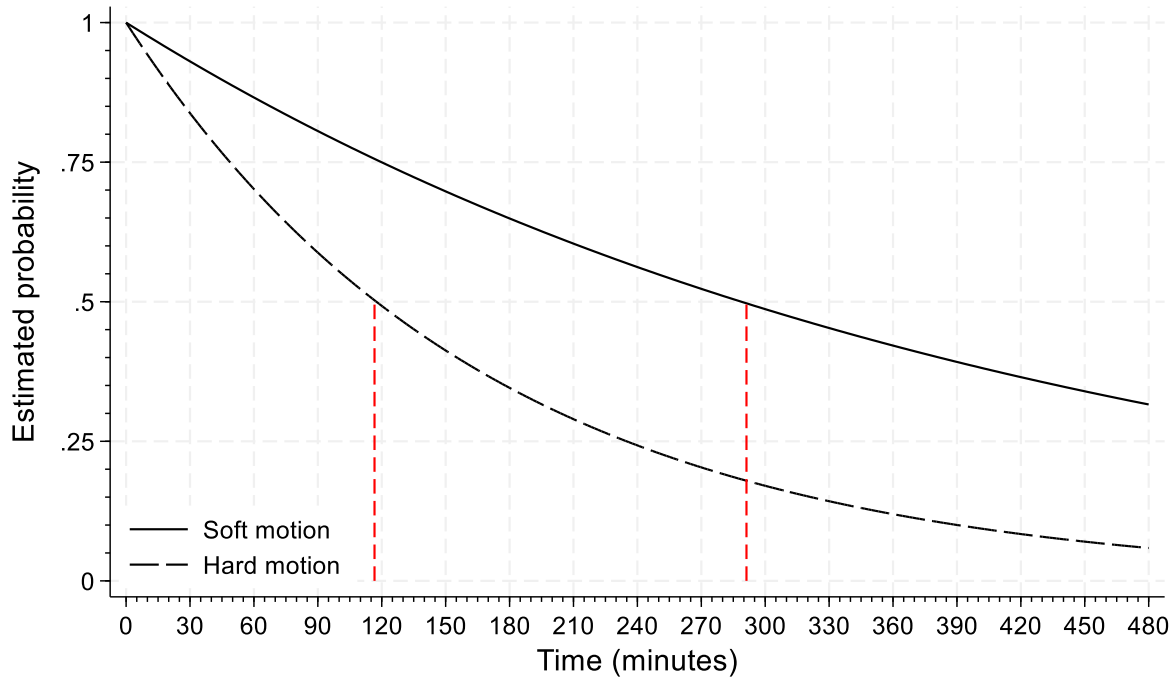

Figure 2. Parametric estimates of the survival function for the motion-sickness studies obtained by assuming an exponential distribution of the time to event. The red dashed lines indicate the estimates of the median time to event for each study.

For the exponential distribution, the survival function is a simple exponential function of λ, i.e., $S(t)=e^{-\lambda t}$. For the "soft motion" study, we get the estimate $S(t)=e^{-0.0024t}$; the plot of the function is presented as the solid black line in Figure 2. Note that the plot covers the values of time t up to an including 480 minutes, to include the estimated mean value of 420.2 minutes. However, as mentioned earlier, the values of the estimated survival function for times larger than 120 minutes are extrapolations beyond the range of the observed data. Thus, their validity very much depends on the assumption of the exponential distribution.

The estimated value of the survival function at, for instance, t=60 is equal to $e^{-0.0024\cdot 60} = 0.866$. This means that the probability that the event will not occur before 60 minutes is equal to about 86.6%. On the other hand, the probability that the event will not occur before 90 minutes is equal to about $e^{-0.0024\cdot 90} = 0.806$, i.e., 80.6%.

For the "hard motion" study, we get $S(t)=e^{-0.0059t}$ (the dashed black line in Figure 2), which leads to $S(60)=e^{-0.0059\cdot 60} = 0.702$ and $S(90)=e^{-0.0059\cdot 90} = 0.588$. This means that the probability that the event will not occur before 60 and 90 minutes is equal to about 70.2% and 58.8%, respectively. These values are smaller than the corresponding estimates (86.6% and 80.6%, respectively) obtained for the "soft motion" study. In fact, this is true for all times t smaller than or equal to 120 minutes, as can be seen from Figure 2. They suggest that the increase of the frequency and the acceleration shortened, on average, the TTE, because the probability that the event will happen after a particular time is estimated to be smaller for the "hard motion" study.

Accelerated failure-time model

The difference in the estimated means, medians, and survival functions suggests that, on average, the TTE for individuals subjected to a lower frequency and acceleration might be longer as compared to a higher frequency and acceleration. However, the difference may be due to chance. Thus, we should

apply a formal significance test. For the exponential distribution, the mean, median, and survival function depend on a single parameter λ. Thus, testing the null hypothesis about the equality of the means, medians, or the survival functions is equivalent to testing the null hypothesis about the equality of the value of λ. However, this is not the case for other distributions which may depend on more than one parameter. For instance, the survival function of the *Weibull distribution* is defined as follows: $S(t)=\exp(-\lambda^{\eta}t^{\eta})$, where λ is called the *scale parameter* and η is called the *shape parameter*. It appears that the exponential distribution is a particular case of the Weibull distribution with $\eta=1$. The median of a Weibull distribution is given by $\lambda\{\ln(2)\}^{1/\eta}$, while the mean is equal to $\Gamma(1+1/\eta)/\lambda$, where $\Gamma(x)$ is the gamma function. Thus, testing the null hypothesis about the equality of, for instance, means for two Weibull distributions implies the need to test the equality of a function of two parameters that have to be estimated.

A useful, general framework to test hypotheses about the mean of a TTE is the *accelerated failure-time* (AFT) model. The model allows evaluating the effect of explanatory variable(s) on the mean of a TTE. In particular, the effect is expressed by multiplying the reference mean by a constant. For instance, for the two motion-sickness studies, the AFT model assumes that

$$\text{(mean of TTE for the “hard motion” study)} = \text{(mean of TTE for the “soft motion” study)}\cdot e^{\beta}$$

where e^{β} is the multiplier expressing the relative change of the mean for the “hard motion” study as compared to the “soft motion” one, and β is a coefficient that has to be estimated from the data. We will refer to the multiplier as the *mean ratio* (MR), because we can transform the equation above as follows:

$$e^{\beta} = \text{(mean of TTE for the “hard motion” study)/(mean of TTE for the “soft motion” study)}.$$

The reason why the MR is expressed as e^{β} is that this form guarantees that the ratio is positive, because it should correspond to a ratio of two non-zero means. Note that, for $\beta<0$, we get $e^{\beta}<1$, which indicates that the mean for the “hard motion” study is smaller than for the “soft motion” (reference) study. This implies acceleration of the events for the “hard motion” study (hence the name of the model). On the other hand, for $\beta>0$, we get $e^{\beta}>1$, i.e., deceleration of the events. In the special case of $\beta=0$ we get $e^{\beta}=1$, i.e., the equality of the means. Thus, $\beta=0$ is the null hypothesis that we would like to test.

If, in order to estimate coefficient β of the AFT model, we make an assumption about the form of the distribution of the TTE, then we talk about a *parametric AFT* model. In that case, β is estimated by using the method of *maximum likelihood*. We are not going to discuss the details of the method here. For the two motion-sickness studies, and assuming that the TTE is exponentially distributed, we get an estimate of $\beta=-0.92$ and $MR = e^{\beta} = e^{-0.92} = 0.4$ (see Table 1). Thus, the MR suggests that the true mean of the TTE for the “hard motion” study was about 60% shorter than for the “soft motion” study. The estimated standard error (SE) of the estimated β is equal to 0.52. Thus, the 95% confidence interval (CI) for the true value of β is equal to $(-0.92-2\cdot0.52, -0.92+2\cdot0.52) = (-1.96, 0.12)$. It includes 0, so we cannot reject the null hypothesis that the means of the TTE for the two studies are equal. By exponentiating the limits of the CI, we obtain the 95% CI for the MR, which is equal to $(e^{-1.96}, e^{0.12}) = (0.14, 1.13)$. As this interval includes 1, we cannot exclude the possibility that the true mean ratio is equal to 1, i.e., that the means of the TTE for the two studies are equal.

Model variant	AFT model		PH model	
	β (95% CI)	MR=e^{β} (95% CI)	θ (95% CI)	HR=e^{θ} (95% CI)
exponential	-0.92 (-1.96, 0.12)	0.40 (0.14, 1.13)	0.92 (-0.12, 1.96)	2.50 (0.89, 7.10)
Weibull	-0.80 (-1.76, 0.16)	0.45 (0.17, 1.17)	0.92 (-0.12, 1.96)	2.50 (0.89, 7.10)
semi-parametric	-0.73 (-1.83, 0.37)	0.48 (0.16, 1.47)	0.90 (-0.14, 1.94)	2.46 (0.87, 6.96)

Table 1. Results of the parametric and semi-parametric models for the motion-sickness studies. AFT: accelerated failure-time; PH: proportional-hazards

Note that, previously, we estimated that the mean values of the TTE for the "soft motion" and "hard motion" study were equal to 420.2 and 168.3 minutes, respectively. The ratio of these values is equal to 168.3/420.2=0.4, i.e., gives the same value as the MR obtained from the AFT model. This is not a coincidence; in this simple case, the correspondence is expected. However, the application of the model offers us a simple means to test the null hypothesis of the equality of the means.

For illustration, we could also apply the AFT model while assuming that the distribution of the TTE is Weibull. In that case, we estimate β=-0.80 and MR = e^{β} = $e^{-0.80}$ = 0.45 (see Table 1). Thus, we get an estimated MR similar to the one obtained for the exponential AFT model. The estimated standard error (SE) of the estimated β is equal to 0.48 and the resulting 95% CI for the true value of β is equal to (-1.76, 0.16). As for the exponential AFT model, the CI includes 0, so we cannot reject the null hypothesis that the means of the TTE for the two studies are equal. The same conclusion is obtained by using the 95% CI for the MR that is equal to ($e^{-1.76}$, $e^{0.16}$) = (0.17, 1.17) and includes the value of 1. For the Weibull model, we also get an estimated value of η, which is equal to 1.15 with the 95% CI equal to (0.76, 1.75). The CI includes the value of 1, what suggests that we could simplify the model and assume the exponential distribution (as we have mentioned earlier, the exponential distribution is a special case of the Weibull distribution for which η=1).

An advantage of the AFT model is that it can be combined with many distributions. Its results are easily interpretable in terms of the increase/reduction of the mean TTE. One important issue is the choice of the parametric distribution. There are several methods that can be used to guide the choice or check if the selected distribution fits the data; we will not review these methods here. A potential, systematic solution to this issue is the use of a *semi-parametric AFT* model (that does not require assumptions about the distribution of the TTE), which will be discussed later in the text.

Proportional hazards model

Proportional hazards (PH) model, also called the *Cox model*, is an often-considered alternative to the AFT model. The popularity of this model is due to its semi-parametric version, which will be discussed later in the text. However, there are parametric versions of the model. Unlike for the AFT model, the PH model can only be applied for selected distributions, which include the exponential distribution and the Weibull distribution.

The PH model is defined on the scale of the *hazard function*. The hazard function, often denoted by $\lambda(t)$, can be interpreted as the instantaneous risk of experiencing the event among individuals who are still at risk (have not had an event and are still exposed to it) at time t. The hazard function is sometimes referred to as the *hazard rate*. Note that is a non-negative function, i.e., $\lambda(t)\geq 0$ for all values of $t\geq 0$. The knowledge of the hazard function completely determines the distribution of the TTE. For instance, the survival function can be written in terms of the hazard function, and *vice versa*.

For the exponential distribution with the survival function $S(t)=e^{-\lambda t}$, the hazard function is constant, i.e., $\lambda(t)=\lambda$, where λ is the parameter characterizing the distribution. For the Weibull distribution with the survival function $S(t)=\exp(-\lambda^{\eta}t^{\eta})$, the hazard function involves a power of the time, i.e., $\lambda(t)=\eta\lambda^{\eta}t^{\eta-1}$.

The form of the PH model is similar to the form of the AFT model, except that the effect of explanatory variable(s) is expressed in terms of multiplying the reference (often called baseline) hazard function. For instance, for the two motion-sickness studies, the PH model assumes that

(hazard rate for the "hard motion" study at t) = (hazard rate for the "soft motion" study at t)$\cdot e^{\theta}$

where e^{θ} is the multiplier expressing the proportional change of the hazard function (hence the name of the model) for the "hard motion" study as compared to the "soft motion" one, and θ is a coefficient that has to be estimated from the data. We will refer to the multiplier as the *hazard ratio* (HR), because we can transform the equation above as follows:

e^{θ} = (hazard rate for the "hard motion" study at t)/(hazard rate for the "soft motion" study at t).

Expressing the HR as e^{θ} guarantees that the ratio is positive, because it should correspond to a ratio of two non-zero functions. Note that, for θ <0, we get $e^{\theta}<1$, which indicates that the hazard rate for the "hard motion" study is smaller (at any time t) than the hazard rate for the "soft motion" (reference) study. On the other hand, for θ>0, we get $e^{\theta}>1$, i.e., an increase of the risk of the event in the "hard motion" study. In the special case of θ=0 we get $e^{\theta}=1$, i.e., the equality of the hazards. Thus, θ=0 is the null hypothesis that we would like to test.

It is worth noting that, in practice, the (true) ratio of two hazard functions may itself be a function of time t. However, the PH model requires that the ratio is constant in time. This is a very strong assumption, which needs to be checked; we will not discuss the methods for checking the PH assumption here.

In case of the parametric PH model, θ is estimated by using the method of maximum likelihood. We are not going to discuss the details of the method here. For the two motion-sickness studies, and assuming that the TTE is exponentially distributed, we estimate θ=0.92 and HR = $e^{\theta} = e^{0.92} = 2.5$ (see Table 1). Thus, the HR suggests that the hazard function for the "hard motion" study is 2.5 times larger than for the "soft motion" study. The estimated standard error (SE) of the estimated θ is equal to 0.52. Thus, the 95% CI for the true value of β is equal to (-0.12, 1.96). It includes 0, so we cannot reject the null hypothesis that the hazard functions of the TTE for the two studies are equal. Consequently, all the characteristics of the distribution of the TTE (e.g., means, medians, survival functions, etc.) are the same. The same conclusion is obtained if we consider the 95% CI for the HR, which is equal to ($e^{-0.12}$, $e^{1.96}$) = (0.89, 7.10), because the interval includes HR=1.

Note that, previously, we estimated the value of the exponential parameter λ to be equal to 0.0024 and 0.0059 for the first and the second study, respectively. From these values we get 0.0059/0.0024=2.46, which is a value very close to the HR obtained from the PH model.

One could also note that the estimated value of θ=0.92 is exactly the negative of the estimated value of β=-0.92 for the AFT model. This is not a coincidence, but the consequence of the fact that, for the exponential distribution, the mean value (1/λ) is the inverse of the hazard rate (λ). Thus, if the hazard rate is modified to λe^{θ}, then the mean is modified to $1/(\lambda e^{\theta})=(1/\lambda)e^{-\theta}=(1/\lambda)e^{\beta}$, so that $\beta=-\theta$.

For illustration, we could also apply the PH model while assuming that the distribution of the TTE is Weibull. In that case, we get, essentially, the same results as for the exponential model: θ=0.92 with SE=0.52 (see Table 1). Thus, we arrive at the same conclusion that we cannot reject the null hypothesis that the hazard functions of the TTE for the two studies are equal.

It is worth noting that the interpretation of the results of the exponential and Weibull PH models is not very intuitive. Taken at face value, the obtained value of HR=2.5 informs us that the instantaneous risk

of the event (hazard rate) for a person subjected to the "hard motion" is 2.5 times higher than for a person subjected to the "soft motion". However, this information does not allow us to conclude anything about, for instance, the difference in the mean or median TTE. In this respect, the result of the AFT model is more direct and easier to understand.

Thus, as compared to the AFT model, the PH model is less intuitive and interpretable. It can be combined with fewer distributions than the AFT model. Moreover, the PH assumption that the ratio of hazard functions is constant in time, is very strong. It has an important, negative consequence: if the model that is applied to a dataset omits an important explanatory variable, the coefficients of the variables included in the model become biased [3]. For randomized clinical trials, this implies that the PH model that includes treatment indicator as the only explanatory variable is likely to underestimate the treatment effect, which may also result in a non-significant result of the test of the effect. The AFT model is not subject to this constraint.

One important advantage of the PH model is that its semi-parametric version has been available since 1970s with accessible software implementation. We will discuss this version later in the text.

Non-parametric methods

It is possible to estimate the survival function non-parametrically. Toward this aim, the most often used method is the *Kaplan-Meier estimator* [4]. Its underlying idea is simple and intuitive: to survive t days, say, first we have got to survive t-1 days, and then the t^{th} day. This implies that we have S(t)=S(t-1)·(probability of no event happening at time t). Consequently, we can estimate the survival function recursively starting from t=0, at which we put S(0)=1, i.e., we assume that we observe TTE for individuals that are event-free at the start of our observation. We will explain the procedure by using the data for the "hard motion" study.

We start with putting S(0)=1, as explained above. To estimate the value of S(1), we have got to estimate the probability that there will be no event at time t=1. We can compute this probability by the ratio of the number of participants that did not experience the event at t=1 relative to the total number of participants that remained under observation (and were, therefore, at risk/exposed to the event) at that time. In our data, all 28 participants remained under observation at the start of the first minute, and none of them experienced the event. Thus, the probability of no event happening at time t=1 can be computed as 28/28=1. Consequently, we get S(1)=S(0)·(28/28)=1·1=1.

For t=2, the situation does not change: 28 participants remain under the observation at the start of the second minute, with none of them experiencing the event. Thus, S(2)=S(1)·(28/28)=1·1=1.

The calculations follow in the same way for t=3 and 4, with S(3)=1 and S(4)=1, respectively.

For t=5, the situation changes: among the 28 participants that remained under the observation at the start of the fifth minute, one experienced the event. This means that the probability of no event happening at time t=5 can be estimated as (28-1)/28=27/28=0.964. As a consequence, we get S(5)=S(4)·(27/28)=1·0.964=0.964.

At t=6, we have got 27 participants remaining under the observation (because we had one event at t=5), with none of them experiencing the event. Thus, S(6)=S(5)·(27/27)= 0.964·1=0.964.

At t=7, the situation changes: we have got 26 participants remaining under the observation, because we "lost" the individual for which we have got the right-censored observation at t=6. However, none of the 26 participants experienced the event. Thus, S(7)=S(6)·(26/26)= 0.964·1=0.964.

The calculations follow in the same way for t=8, 9, and 10, with S(8)=S(9)=S(10)=1.

At t=11, among the 26 participants that remained under the observation at the start of the 10th minute, two experienced the event. This means that the probability of no event happening at time t=11 can be estimated as (26-2)/26=24/26=0.923. As a consequence, we get S(11) = S(10)·(24/26) = 0.964·0.923 = 0.890.

At this point, the idea of the estimating procedure should be clear. A couple of comments are worth making. First, the estimated value of the survival function changed only at the times at which an event was observed. Thus, for instance, while the "hard motion" study included 28 participants, the calculations have got to be conducted only for 11 values of time, because there were only 11 distinct uncensored times observed in the study. Second, the estimated value of the survival function did not change at t=6, the time at which only a right-censored observation was recorded. However, at the subsequent time(s), the number of participants remaining at observation was reduced. Thus, the censored observation was included in all the calculations for the preceding times. In this way, the partial information conveyed by the censored observation (that the participant did not experience event until t=6) was used in our estimation procedure. This illustrates how survival-analysis methods take into account the censored observations.

For completeness, the estimated values of the survival function for both studies are provided in Table 2.

"Soft motion" study			"Hard motion" study		
time	K-M estimator	exponential	Time	K-M estimator	exponential
30	0.952	0.930	5	0.964	0.971
50	0.905	0.887	11	0.890	0.937
51	0.854	0.885	13	0.853	0.926
82	0.801	0.821	24	0.816	0.868
92	0.748	0.802	63	0.779	0.689
			65	0.742	0.681
			69	0.667	0.665
			79	0.630	0.627
			82	0.556	0.616
			102	0.519	0.548
			115	0.482	0.507

Table 2. Non-parametric estimates of the survival function for the motion-sickness studies obtained by using the Kaplan-Meier estimator. For comparison, the parametric estimates obtained by assuming the exponential distribution are also provided. K-M: Kaplan-Meier.

The plot of the Kaplan-Meier estimator of the survival function is often called a *survival curve*. Figure 3 presents the survival curves for the motion-sickness studies. Both curves have a stepped shape, which reflects the comment just made that the Kaplan-Meier estimator changes its value only at the times at which events are recorded. The black tick marks with numbers on the curves indicate the number of the right-censored observations recorded at the particular time. It is worth observing that both curves are plotted for values of time until t=120 minutes. This is because the non-parametric estimator, unlike the parametric estimator, cannot be extrapolated beyond the range of times observed in the data.

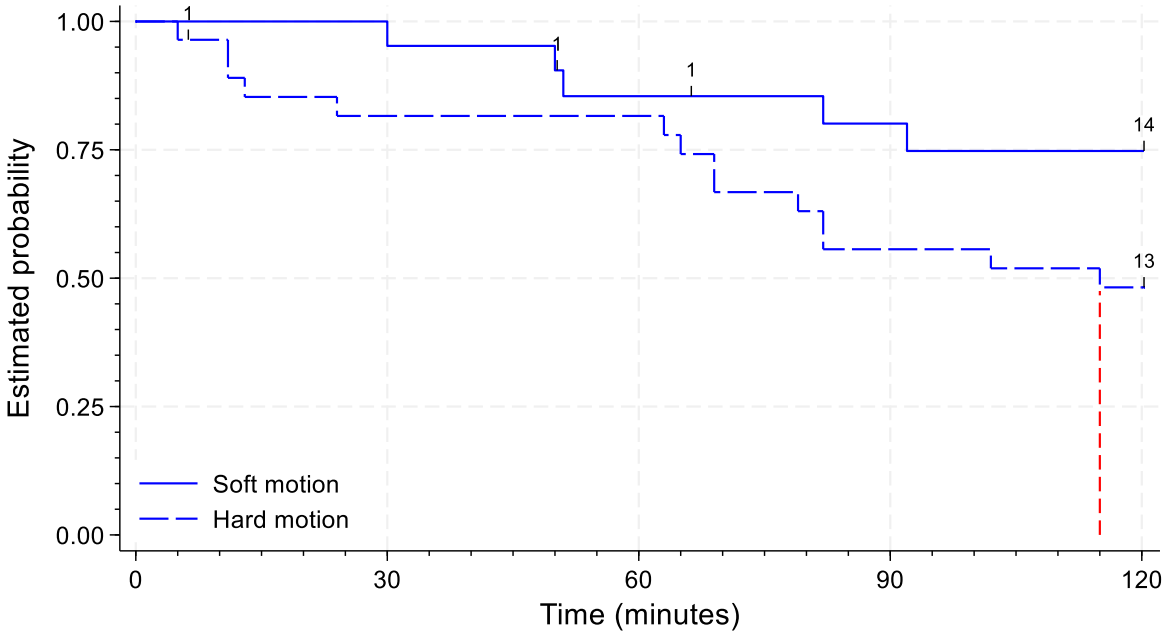

Figure 3. Non-parametric estimates of the survival function for the motion-sickness studies obtained by using the Kaplan-Meier estimator. The red dashed line indicates the estimate of the median time to event for the "hard motion" study.

For the "soft motion" study, the Kaplan-Meier estimator indicates that the probability that the event will not occur before 60 and 90 minutes is equal to about 85.0% and 80.0%, respectively. For the "hard motion" study, the estimates are equal to 81.0% and 56.0%, respectively. These non-parametric estimates agree fairly well with the estimates obtained by assuming that the TTE is exponentially distributed (86.6% and 80.6% for the "soft motion" study, respectively, and 70.2% and 58.8% for the "hard motion" study, respectively). Figure 4 shows that, indeed, the Kaplan-Meier estimates agree reasonably well, within the range of the observed times, with the estimates obtained by using the exponential assumption (see also Table 2). Note that the figure clearly illustrates the fact that any inference regarding the distribution of the data beyond 120 minutes requires an extrapolation beyond the observed data range and depends on the validity of the parametric assumption. Despite the agreement between the non-parametric and parametric estimates for times up to 120 minutes, it is not guaranteed that the parametric assumption would be appropriate for longer times.

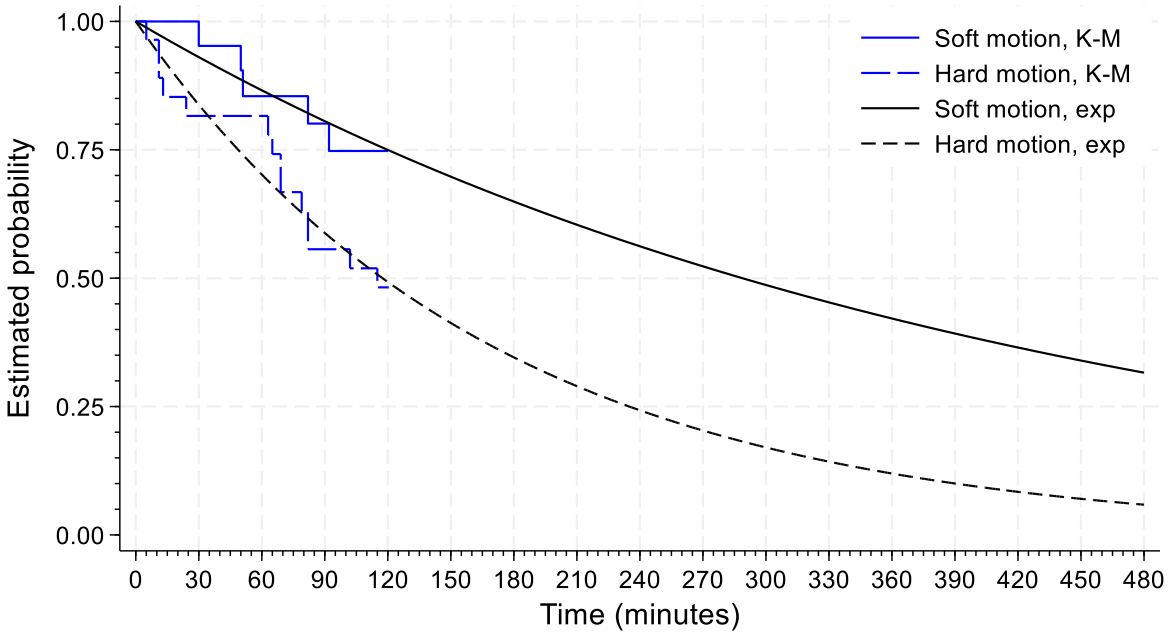

Figure 4. Comparison of the non-parametric and parametric estimates of the survival function for the motion-sickness studies.

The Kaplan-Meier estimator can be used to obtain a non-parametric estimate of the median TTE. Towards this aim, one may use the time, at which the estimator "drops" below 50%. For instance, Figure 3 shows that, for the "hard motion" study, the median TTE can be estimated to be equal to 115 minutes (indicated by the vertical red dashed line). The value agrees very well with the estimate of 116.6 minutes obtained by assuming that the distribution of TTE is exponential. On the other hand, for the "hard motion" study, we cannot estimate the median, because the survival curve never reaches values below 50%. This illustrates the difficulty in the non-parametric estimation of the median.

The Kaplan-Meier estimator can also be used to obtain a non-parametric estimate of the mean TTE by computing the area under the survival curve. However, the estimate is only valid if the survival curve reaches the value of 0. As seen in Figure 3, this is the case of neither of the motion-sickness studies. As a result, we cannot obtain the non-parametric estimate of the mean TTE based on the data from the two studies.

The Kaplan-Meier estimates and the exponential-distribution-based estimates (see Figure 4) suggest that the true survival functions for the two studies may be different. Thus, we might want to formally test the null hypothesis that the functions are the same. Toward this aim, we may use the *logrank test*. This is a non-parametric test. The underlying idea is to compare the observed number of the events with the number of events that would be expected under that null hypothesis, i.e., when assuming that the two survival functions are the same.

To get an insight into the idea and construction of the test, we will illustrate the calculations necessary to conduct it by using the data from the two motion-sickness studies.

First, we have to construct an ordered list of all uncensored times recorded in any of the studies. The list looks as follows: 5, 11, 13, 24, 30, 50, 51, 63, 65, 69, 79, 82, 92, 102, 115. Calculations are conducted for each time from the list, starting from the first one.

At t=5 minutes, there were, in total, 49 participants under observation in both studies, with 21 (42.9%) in the "soft motion" study and 28 (57.1%) in the "hard motion" study. At that minute there was, in total, one event observed (in the "hard motion" study). Under the null hypothesis, there should be no difference in the probability of observing an event for both studies. Thus, we would expect that the events should be distributed between the studies according to the fraction of participants that were exposed to the event in each study. Consequently, at t=5, we would expect 1·0.429 = 0.429 events in the "soft motion" study and 1·0.571 = 0.571 in the "hard motion" one. Note that we allow the expected number of events not to be an integer.

At t=11 minutes, there were, in total, 47 participants under observation in both studies (in the "hard motion" study, one individual had an event at t=5 and one was "lost" from observation at t=6), with 21 (44.7%) in the "soft motion" study and 26 (55.3%) in the "hard motion" study. At that minute there were, in total, two events observed (both in the "hard motion" study). Under the null hypothesis, we would expect 2·0.447= 0.894 events in the "soft motion" study and 2·0.553 = 1.106 in the "hard motion" one.

At t=13 minutes, there were, in total, 45 participants under observation in both studies (as compared to t=11, two individuals had events at t=11 in the "hard motion" study), with 21 (46.7%) in the "soft motion" study and 24 (53.3%) in the "hard motion" study. At that minute there was, in total, one event observed (in the "hard motion" study). Thus, under the null hypothesis, we would expect 1·0.467= 0.467 events in the "soft motion" study and 1·0.533 = 0.533 in the "hard motion" one.

At this point, the idea of the calculations should be clear. After completing them for all the times from the list (see Table 3), we sum the numbers of expected events for each study. As a result, we obtain 8.86 expected events for the "soft motion" study and 10.14 events for the "hard motion" study. Note that their sum (8.86+10.14=19) is equal to the total number of events (5+14=19) observed in both studies. By definition, this should be the case; this is a useful check of the correctness of the calculations.

It is worth noting that the, in the "soft motion" study, the number of expected events (8.86) is larger than the number (5) of the observed ones, with the opposite pattern occurring in the "hard motion" study. This suggests that, in the "soft motion" study, we observed fewer events than would be expected under the null hypothesis, which might be taken as a signal that the risk of the event in that study would be smaller.

However, the differences between the observed and expected numbers of events may be due to the play of chance. To formally compare them, we construct the following test statistic:

$$(8.86-5)^2/8.86 + (10.14-14)^2/10.14=3.15.$$

Note that, for each group, we take the squared difference (which allows disregarding the sign of the difference) and make it relative to the number of the expected events (to make it comparable). Now, the question is, whether the obtained value of the statistic (3.15) is large or small? It appears that, assuming the null hypothesis, we would observe, for the constructed statistic, values as large as 3.15 with probability p=0.076. This is the *p-value* of our test; it was computed by using the chi-squared distribution with 1 degree of freedom. Given that the p-value is larger than the often-used significance level of 0.05, we cannot conclude that the result of the logrank test is statistically significant at that significance level. Thus, we cannot reject the null hypothesis that the survival functions for the two studies may be the same, despite the differences observed in Figure 3.

Both studies			"Soft motion" study			"Hard motion" study		
time	exposed	events (obs)	exposed	events (obs)	events (exp)	exposed	events (obs)	events (exp)
5	49	1	21	0	0.429	28	1	0.571
11	47	2	21	0	0.894	26	2	1.106
13	45	1	21	0	0.467	24	1	0.533
24	44	1	21	0	0.477	23	1	0.523
30	43	1	21	1	0.488	22	0	0.512
50	42	1	20	1	0.476	22	0	0.524
51	40	1	18	1	0.450	22	0	0.550
63	39	1	17	0	0.436	22	1	0.564
65	38	1	17	0	0.447	21	1	0.553
69	36	2	16	0	0.889	20	2	1.111
79	34	1	16	0	0.471	18	1	0.529
82	33	3	16	1	1.455	17	2	1.545
92	30	1	15	1	0.500	15	0	0.500
102	29	1	14	0	0.483	15	1	0.517
115	28	1	14	0	0.500	14	1	0.550

Table 3. Results of the computation of the expected number events (given in the sixth and nineth column) for the motion-sickness studies. For the explanation of the details of the computation, see the text.

A couple of comments are worth making here. First, the procedure presented above illustrates a simplified calculation of the test statistic. A more precise (though numerically more cumbersome) procedure leads to the value of 3.12 and p=0.073, without any difference for the conclusion. Second, the procedure can be extended to comparison of three or more survival functions by using the appropriate fractions of individuals remaining under the observation at each event time. In that case, the p-value of the test has to computed by using the chi-squared distribution with the number of degrees of freedom equal to the number of compared survival functions less one.

Semi-parametric AFT and PH models

As it has been mentioned earlier, the AFT and PH models can be used in a semi-parametric form. This means that they can be applied without making assumptions about the distribution of the TTE. Note that the models still require the estimation of the coefficients β (AFT) and θ (PH). For this reason, the models cannot be termed non-parametric.

The main advantage of avoiding the need to make assumptions about the distribution of the TTE is the broader applicability of the models. However, the models still require that their assumptions about the proportional change of the mean (AFT) or hazard function (PH) are met. As it has been mentioned earlier, especially the PH assumption is a strong one.

The *semi-parametric PH model* was developed in 1972 [5]. Almost immediately, software allowing the estimation of the coefficients (θ) of the model became available. This resulted in an enormous popularity of the semi-parametric PH model. For instance, treatment effects in clinical trials in oncology are almost exclusively reported in terms of the HR.

The *semi-parametric AFT model* was proposed in 1981 [6]. However, the estimation of the coefficients (β) of the model is numerically complicated and no software implementation was readily available. This difficulty hampered the use of the model. Thus, for several decades, mainly the parametric version

of the AFT model has been available for practical applications. This was the major obstacle for a widespread use of the AFT model in the context of, for instance, cancer clinical trials.

In recent years, however, important developments have taken place regarding the numerical tools allowing the estimation of the coefficients of the semi-parametric AFT model [7]. Thus, the model is available for practical applications.

For the data obtained in the two motion-sickness studies, the semi-parametric AFT model yields an estimate of β=-0.73 and MR = $e^{\beta} = e^{-0.73} = 0.48$ (see Table 1). The MR value is somewhat closer to 1 as compared to the estimates obtained for the exponential and Weibull models. The estimated SE of the estimated β is equal to 0.55. As a result, the 95% CI for the true value of β is equal to (-1.83, 0.37). As for the exponential and Weibull AFT models, it includes 0, so we cannot reject the null hypothesis that the means of the TTE for the two studies are equal. The same conclusion is obtained by using the 95% CI for the MR, which is equal to $(e^{-1.83}, e^{0.37}) = (0.16, 1.47)$ and includes MR=1.

The semi-parametric PH model provides the estimates of θ=0.90 and HR = $e^{\theta} = e^{0.9} = 2.46$ (see Table 1). The HR is comparable to the estimates obtained for the exponential and Weibull PH models. The estimated SE of the estimated θ is equal to 0.52 and leads to the 95% CI for the true value of β equal to (-0.14, 1.94). As for the exponential and Weibull PH models, it includes 0, so we cannot reject the null hypothesis that the hazard functions of the TTE for the two studies are equal. The same conclusion is obtained by using the 95% CI for the HR, which is equal to $(e^{-0.14}, e^{1.94}) = (0.87, 6.96)$ and includes HR=1.

For the two motion-sickness studies, the results of the parametric and semi-parametric variants of the AFT and PH models are similar. As it is seen in Figure 4, the exponential-distribution-based estimates of the survival function agree quite well with the result of the non-parametric Kaplan-Meier estimator. Thus, it seems that, in the case of data collected in the two studies, the assumption that the TTE has an exponential (or Weibull) distribution may actually be appropriate. Taking this into account, the agreement between the results obtained by the parametric and semi-parametric models should not be surprising.

It is worth noting that the (parametric and semi-parametric) AFT and PH models can easily be extended to allow for effects of more than one explanatory variable. For instance, if the motion-sickness studies provided information about the sex of participants, one could consider, for instance, the following AFT model (a similar extension would work for the PH model):

(TTE-mean for females and “hard motion”) = (TTE-mean for females and “soft motion”)$\cdot e^{\beta}$

(TTE-mean for males and “soft motion”) = (TTE-mean for females and “soft motion”)$\cdot e^{\delta}$

(TTE-mean for males and “hard motion”) = (TTE-mean for females and “soft motion”)$\cdot e^{\beta} \cdot e^{\delta}$

In this model, the multiplier e^{β} describes the change of the mean TTE for “hard motion” as compared to the “soft motion”, while e^{δ} describes the change of the mean TTE for males. Note that, by using the product $e^{\beta} \cdot e^{\delta}$, we assume that the change of the mean due to the nature of motion is independent of the change due to the sex. If we wanted to make the changes to depend on each other, we could introduce the *interaction effect*, e^{τ} say, and specify that

(TTE-mean for males and “hard motion”) = (TTE-mean for females and “soft motion”)$\cdot e^{\beta} \cdot e^{\delta} \cdot e^{\tau}$

In this model, the relative effect of “hard motion” for males is equal to $e^{\beta} \cdot e^{\tau}$, i.e., is additionally modified by e^{τ} as compared to the effect of “hard motion” for females, which is equal to e^{β}.

Discussion

When the collected data include censored observations, classical methods of statistical analysis such as the sample mean, sample median, linear regression etc. cannot be applied. This is because they, in general, produce biased results. Sometimes censored observations are dealt with by using some simplistic "imputation" techniques in an attempt to apply the classical methods. For instance, measurements below the lower limit of detection (i.e., left-censored observations) are replaced by the half of the limit and, subsequently, the sample mean is used to estimate the (unknown) true mean value. Such approaches are best to be avoided, because they do not remove the bias. Instead, the data should be analyzed by using survival-analysis methods, which properly take into account the presence of censored observations.

In this article, parametric and non-parametric methods of survival analysis have been discussed. Both approaches are applicable in dentistry and orthodontics. For instance, Kuthy et al. (2014) applied the Weibull-distribution-based parametric PH model to analyze the data collected in their retrospective study.

The choice between the parametric and non-parametric approaches depends on the data at hand. Parametric methods yield, in general, more precise estimates and more powerful tests of statistical significance, thus they may be useful for smaller sample sizes. However, the results crucially depend on the validity of the parametric assumption. Hence, the validity should be carefully checked. Non-parametric methods are more robust in that respect, but they require larger sample sizes. Note that, in case of data with censored observations, the information comes primarily from the observed event times, i.e., uncensored observations. Thus, it is not the total sample size that matters, but the number of observed events.

References

1. Kuthy RA, Jones M, Kavand G, Momany E, Askelson N, Chi D, Wehby G, Damiano P. Time until first dental caries for young children first seen in Federally Qualified Health Centers: a retrospective cohort study. Community Dentistry and Oral Epidemiology 2014; 42:300–310.
2. Burns KC. Motion sickness incidence: Distribution of time to first emesis and comparison of some complex motion conditions. Aviation, Space, and Environmental Medicine 1984; 55:521-527.
3. Lin NX, Logan S, Henley WE. Bias and sensitivity analysis when estimating treatment effects from the Cox model with omitted covariates. Biometrics 2013; 69: 850-860.
4. Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 1958; 53:457-481.
5. Cox DR. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society Series B Statistical Methodology 1972; 34:187-220.
6. Louis TA. Nonparametric analysis of an accelerated failure time model. Biometrika 1981; 68: 381-390.
7. Burzykowski T. Semi-parametric accelerated failure-time model: A useful alternative to the proportional-hazards model in cancer clinical trials. Pharmaceutical Statistics 2022; 21:292-308.