University of Liège
Faculty of Sciences
Department of Mathematics
Academic year 2017 - 2018

Simulation-based comparative performance of survival methods in the case of non-proportional hazards.

Supervisors
Anne-Françoise Donneau
Michal Kicenski

Master thesis written by
Emeline Thunus
presented for the Master's degree
in Mathematical Sciences with a
focus on statistics.

First and foremost, I have to thank my supervisors, Ms. Donneau,
Mr. Kicinski and for having offered me this topic,
giving me the opportunity to apply mathematics in practice.
Without their assistance and valuable advice and suggestions
in every step throughout the process, this master thesis
would have never been accomplished.
Thank you also for their availability
throughout this work
and their support.

Then, I would like to thank the EORTC
for supplying the dataset required,
as well as Mr. Peron and Mr. Ozenne
for their availability to solve
certain issues in the simulation study.
I would also like to thank Ms. Colette
and Ms. Mauer for their valuable advice.

Moreover, I thank Ms. Brereton
for taking time and having the patience
to correct my English.
Her advice has helped me
to improve my English.

Most importantly, none of this could have happened
without my family, my friends and Nicolas.
They have always had an open ear for my worries,
have made me feel better in the moments of despondency,
shared my happiness about success
and believed in me over the duration
of my entire studies.

Contents

List of symbols

AWLR	Adaptive Weighted Logrank test
BT0	Generalised pairwise comparison with threshold 0
BT1	Generalised pairwise comparison with threshold 1
Ctr	Control group
$f(t)$	Density function
$G01$	Weighted logrank test with $(\rho, \gamma) = (0, 1)$
$G10$	Weighted logrank test with $(\rho, \gamma) = (1, 0)$
$\mathbb{I}(.)$	Indicator function
LR	Logrank test
RMST	Restricted Mean Survival Time
$S(t)$	Survival function
Trt	Treatment Group
V_i	Weighted Kaplan-Meier test
BFV_i	Weighted Kaplan-Meier test using bona fide p-value
WLR	Weighted logrank statistics
$\lambda(t)$	Hazard function
$\Lambda(t)$	Cumulative hazard function
μ_τ	Restricted mean survival time
Δ	Difference in restricted mean survival time (RMSTD)
$\overline{\Delta}$	Relative difference in restricted mean survival time

- $\Delta^\star$ .... Proportion in favour of treatment
- $\perp\!\!\!\perp$ .... Independence between two random variables

Introduction

As part of the search for a treatment for cancer, statistical studies such as survival analyses can be useful to evaluate the effect of two treatments. During research on a new treatment, a comparative analysis between treatment effects on survival is investigated. The logrank test is the most popular choice when it comes to comparing treatment efficiency in cancer research. It has been well-established that it is the most powerful non-parametric test to compare survival functions when the hazards are proportional over time. However, in the case of the assumption of proportional hazard being in doubt, the logrank test may no longer be optimal. The proportional hazard assumption can easily be verified by using the Kaplan-Meier curve. Formal way, we can use Grambsh and Therneau's approach [1] to indentify non-proportionality of a case. A number of statistical approaches have been investigated to compare treatment effects on survival in the event of the proportional hazard assumption being violated but they are still underused and not very well-known. This lack of application can be explained by the fact that it is currently unclear which approach should be used according to the design and analysis studies when comparing cancer treatments. Nonetheless, Li et al. [2] and Callegaro and Spiessens [3] compare some methods where survival functions cross over time.

In this Master's thesis, a particular focus is given to immunotherapy treatment. Immunotherapy, explained in [4], consists of using the immune system to fight cancer both directly and indirectly. In order to understand what immunotherapy is, we need to become familiar with how the immune system works. The immune system is composed of organs, cells and substances. Its function is to protect and defend the body against foreign agents such as bacteria and viruses. Furthermore, the immune system can differentiate between the healthy cells of the body and cancer cells, and can also fight them. The immune system works as a memory. Indeed, it memorizes a trace of all substances that can be found in the body. Thereby, all new substances which have not been memorized antigens cause a reaction of the immune system. Germs such as viruses and bacteria can detect these antigens. Cancer cells work like antigens but unlike the latter, the immune system has a hard time fighting the former naturally. Several reasons make this fight difficult. Indeed, in many cases, the cancer cells do not differ enough from healthy cells. In that case, the immune system does not consider the cancer cells to be a threat. In other cases, cancer cells are considered foreign cells but the action taken by the immune system to fight the cancer is not powerful enough. Moreover, cancer cells can thwart immune action, releasing a substance which neutralizes the immune system.

To fix these problems, different approaches of immunotherapy have been designed in order to support the immune system. Immunotherapy treatment methods lead to a delayed clinical effect. The issue of how to compare treatment effects when the hazards are likely non-proportional is of major interest for contemporary cancer research due to the ground-breaking achievements of immunotherapy, which is typically characterized by a delayed treatment effect. A delayed treatment effect means that at the beginning of the study there are no differences between the survival curves of treatment and control groups. Nonetheless, at a certain time, a treatment effect appears which is illustrated by a split between survival curves. The aim of the present master thesis was to compare the classical logrank test and the methods designed to deal with non-proportional hazards.

The master thesis is structured as follows: In the first chapter, we briefly introduce the background of survival analysis. We define what survival analysis is, censoring, survival function and hazard function are. We demonstrate the Kaplan-Meier estimator of the survival function. Then we describe the aim of a test comparing survival curves and finally explain the proportional hazard assumption. Statistical methods, such as logrank, restricted mean survival time, generalised pairwise comparison, weighted logrank, adaptive weighted logrank and weighted Kaplan-Meier test are described in the second chapter. Then we use three types of simulations to compare all methods according to the type I-error. Finally, all methods are illustrated on the real EORTC 18991 trial outcome data to compare adjuvant therapy with pegylated-interfon-α-2b with observation in stage III melanoma patients.

Chapter 1

Survival analysis

The aim of survival analysis is to model time until an event occurs. As an example, when the event of interest is "death", the outcome variable will be the elapsed time between the beginning of the follow-up of an individual until their death. Within the context of cancer studies, the main events of interest are either death, or relapse from remission or from toxicity on vital organs.

In survival analysis, we usually refer to the time until an event occurs as survival time. However, survival data can be censored; this occurs when the individual does not show the event of interest. In that situation, we have some information about individual survival time; however the exact survival time is unknown. For example, the event is not observed before the end of the study, or the individual withdraws from the study.

Classic endpoints in cancer clinical trials are usually defined in terms of survival. For that purpose, the main objective is to compare the time until one such event occurs between two groups of patients, namely a control and a treatment group.

In this chapter, based on [5], we will first describe some background of the survival analysis that includes functions, the relationships between them. These functions allow us to describe the process studied and the probability of an event occurring at a certain time. Then, we will show how to estimate those new functions and illustrate on an example. Finally, we will use a classic statistical test to compare two survival distributions. The first one will be a treatment group and the second one a control group. They will be respectively referred to as "Trt" and "Ctr".

1.1 Background

Consider a non-negative continuous random variable T which takes the value $t_1, ..., t_n$ where $n \in \mathbb{N}_0 \cup \{\infty\}$ with probability density function $f(t)$. This value represents the time which has passed from the time of recruitment in the study to the completed event. Indeed, during a study, not all patients enter the study at the same time. The example 1.1.1.1 illustrates the time from recruitment in the study until the considered event occurs. In this sub-section, we describe the behaviour of the random variable T in order to use it in the survival analysis.

1.1.1 Censoring

When studying time-to-event data, the lack of information about the event can be a limiting factor. Indeed, due to the fact that the duration of the study is limited, it is possible that some events do not occur before the end of the study. In such situations, times are considered as *right censored* because we cannot predict the survival time for these individuals. The only information that we have is that the survival time is greater than the end of the study. Another situation of *right censoring* occurs when the patient withdraws from the study; then, we only have a lower bound for his or her survival time. The survival time T is included in the interval $[\mathrm{C}, +\infty[$ when the time is *right censored.* There exist other types of censoring such as left censored and interval censoring, but they will not be explained in this master thesis.

Regarding the time-to-event, we can have different possibilities:

1. An individual completes the event before the end of the study. In this case, the value of T is completely determined.

2. An individual does not complete the event before the end of study or the individual withdraws from the study. In this case, we know that the individual stay alive until C.

Let δ denote a dichotomous variable indicating event occurrence or censoring. Each individual can then be characterized with a pair of random variables (X, δ) where:

$$X = \begin{cases} T & \text{when uncensored} \\ C & \text{when censored} \end{cases}$$

$$\delta = \begin{cases} 1 & \text{when uncensored} \\ 0 & \text{when censored} \end{cases}$$

Example 1.1.1.1. Let us consider the example of a clinical trial to relieve the symptoms of a chronic medical condition. Recruitment of eligible patient into the trial started on 1^{st} January 2016. Recruitment of eligible patients into the trial ceases on 1^{st} January 2018. The variable of interest T is the time between treatment and relapse. The Figure 1.1 illustrates the elapsed time between recruitment and observed or censored time. Censored observations are represented by red dots. Red dots occur when the survival lines are truncated by the red line. This truncation shows that the survival time of these patients continues after the end of the study.

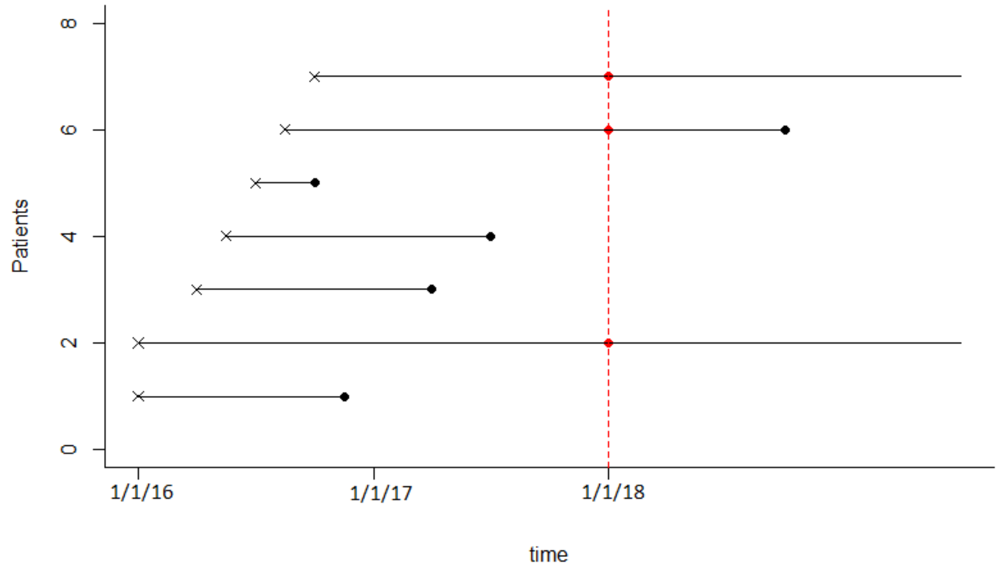


Figure 1.1: Recruitment and censoring

1.1.2 Survival function

Definition 1.1.1. The *survival function* of T provides the probability that the event of interest has not occurred during $[0, t]$:

$$S : [0, +\infty[\to [0, 1] : t \longmapsto S(t) = \mathbb{P}(T > t). \tag{1.1}$$

In the continuous case,

$$S(t) = \int_t^\infty f(t)\mathrm{d}t.$$

From this equality, we get

$$f(t) = -\frac{d}{dt}S(t).$$

From definition 1.1.1, it can be seen that the survival function is a decreasing function taking values with $S(0) = 1$ and $S(t) \to 0$ when $t \to +\infty$.

1.1.3 Hazard function

Definition 1.1.2. The *hazard function* for a random variable T is defined as

$$\lambda : [0,+\infty[\rightarrow [0,1] : t \longmapsto \lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t+\Delta t | T \geq t)}{\Delta t}.$$

The numerator can be thought of as the expected number of future events (failure) in the time interval $[t, t+\Delta t[$.

The quantity $\lambda(t)\Delta t$ may be considered as an approximate conditional probability of an event in the interval $[t, t+\Delta t[$. However, the hazard function is not a probability but the ratio $\dfrac{\mathbb{P}(t \leq T < t+\delta t | T \geq t)}{\Delta t}$ can be interpreted as the event (failure) rate over the time interval $[t, t+\Delta t[$.

The hazard function illustrates how likely the event is to occur in a small interval containing a particular time t, given that it has not occurred before then.

Furthermore, we have the following equality:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

Indeed, when applying the conditional probabilities formula,

$$\begin{aligned} \mathbb{P}(T \leqslant t+\Delta t \mid T > t) &= \frac{\mathbb{P}(T \leqslant t+\Delta t \cap T > t)}{\mathbb{P}(T > t)}, \\ &= \frac{\mathbb{P}(t < T \leqslant t+\Delta t)}{\mathbb{P}(T > t)}, \\ &= \frac{F_T(t+\Delta t) - F_T(t)}{S_T(t)}. \end{aligned}$$

So,

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t \; S(t)}, \\ &= \frac{1}{(t)} \frac{d}{dt} F(t), \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

Then, $\lambda(t) = -\dfrac{d}{dt} ln[S(t)]$.

Definition 1.1.3. The *cumulative hazard function* is defined by

$$\Lambda(t) = \int_0^t h(u)\, \mathrm{d}u.$$

This function describes the "total exposure to risk" for a survivor up to $T = t$.

There is a relationship between $\Lambda(.)$ and $S(.)$:

$$\Lambda(t) = -log(S(t)) \Longleftrightarrow S(t) = exp(-\Lambda(t)).$$

Table 1 summarizes the various relationships that exist between the previously defined functions.

Table 1.1: Relationship between density function,
survival function and hazard function

	$f(t)$	$S(t)$	$\lambda(t)$
$f(t)$		$-\frac{d}{dt}S(t)$	$\lambda(t)\exp\left(-\int^t \lambda(s)\,\mathrm{d}s\right)$
$S(t)$	$\int_t^\infty f(s)\,\mathrm{d}s$		$\exp(-\int_0^t \lambda(s)\,\mathrm{d}s)$
$h(t)$	$\frac{f(t)}{\int_0^\infty f(s)\,\mathrm{d}s}$	$-\frac{d}{dt}log(S(t))$	

1.1.4 Description of observations in terms of failure

At each time t_i , a certain proportion of individuals complete the event of interest or are censored. Let d_i denote the number of events (e.g. number of deaths) and c_i the number of censored events at time t_i. The number of individuals at risk at time t_i is denoted by n_i and is equal to the number of individuals who will die or be censored,

$$n_i = \sum_{t=i}^{n}(d_t + c_t).$$

1.2 Estimation

Certain functions describing the behaviour of the random variable T were introduced into 1.1.2. However, in practice, we do not know the distribution of this variable. We only have a sample of the observations to form the base to estimate the different functions.

Hereafter, we describe one way of estimating survival function. We will use the maximum likelihood to obtain the *number of failures* and then apply the result to obtain an estimator for the survival function $S(t)$, known as the Kaplan-Meier estimator.

1.2.1 Discrete survival distribution

Let $0 < t'_1 < t'_2 < ... < t'_m < \infty$ be the distinct observed failure times in sample $\{t_1, ..., t_n\}$ as it is possible to observe more than one failure at the same time, $m \leq n$. We denote $t'_0 = 0$ and $t'_{m+1} = \infty$.

The probability function can be written as

$$f(t'_i) = \mathbb{P}(T = t'_i), \quad i = 1, ..., m, \tag{1.2}$$

the survival function as

$$S(t) = 1 - \sum_{j:t'_j<t} f(t'_j) = \sum_{j:t'_j>t} f(t'_j), \text{ with } t \in [t'_i, t'_{i+1}[\text{ and } i = 0, ..., m, \tag{1.3}$$

the hazard function as

$$\begin{aligned}\lambda(t'_i) &= \mathbb{P}(T = t'_i | T \geq t'_i), \\ &= \frac{\mathbb{P}(T = t'_i)}{\mathbb{P}(T \geq t'_i)}, \\ &= \frac{\mathbb{P}(T = t'_i)}{\mathbb{P}(T > t'_{i-1})}\end{aligned}$$

then,

$$\lambda(t'_i) = \frac{f(t'_i)}{S(t'_{i-1})}. \tag{1.4}$$

There are constraints from the definition of the survival function such as $S(t'_0) = 1$ and from equations (1.3) and (1.4), we obtain

$$\begin{aligned}S(t'_i) &= S(t'_{i-1}) - f(t'_i), \\ &= S(t'_{i-1}) - \lambda(t'_i)S(t'_{i-1}).\end{aligned}$$

Then,

$$S(t'_i) = S(t'_{i-1})(1 - \lambda(t'_i)). \tag{1.5}$$

Hence,

$$S(t'_i) = \prod_{j=1}^{i}(1 - \lambda(t'_j)) \text{ and } f(t'_i) = \lambda(t'_i)\prod_{j=1}^{i-1}(1 - \lambda(t'_j)). \tag{1.6}$$

In the following sections, for clarity and ease of reading, we denote $S(t'_i)$ as S_i, $f(t'_i)$ as f_i and $\lambda(t'_i)$ as λ_i .

Let d_i be the number of failures observed at time t'_i $(i = 1, ..., m)$, then $\sum_i^m d_i = d_+$.

Let c_i $\quad (i = 1, ..., m)$ be the number of censored observations with censoring times between t'_i and t'_{i+1}.

1.2.2 Maximum discrete likelihood estimation

Maximum likelihood estimation is a general method for parameter estimation. The maximum likelihood estimator $\widehat{\theta}$ maximises the likelihood $L(\theta)$ which is simply the joint probability(density) of the observed data treated as a function of the unknown θ.

Assuming independent and identically distributed observations with no censoring, the likelihood is given by

$$L(\theta) = \prod_{i=1}^{n} f(t_i, \theta),$$

as the joint probability density function of independent variables is just the product of their individual marginal probability density functions.

Now, consider the independent and identically distributed observations with right censoring indicators $\delta_1, ..., \delta_n$. For a right-censored observation, t_i is not an observed value but we know an interval $[c_i, +\infty[$ where the observation is included. Hence, the appropriate contribution to the likelihood for a censored t_i is $S(t_i)$. So, the likelihood is given by

$$L(\theta) = \prod_{i:\delta_i=1} f(t_i, \theta) \prod_{i:\delta_i=0} S(t_i, \theta).$$

Then, with the following property, we obtain an estimator for discrete hazard function at time t_i.

Property 1.2.2.1. *The discrete hazard maximum likelihood estimator is given by*

$$\widehat{\lambda_i} = \frac{d_i}{r_i} \text{ for } i = 1, ..., m,$$

where $r_i = \sum_{j=i}^{m}(d_j + c_j)$ is called the number at risk (of failure) at t'_i.

Proof. We can reformulate the likelihood with the notation introduced previously as follows

$$L = \prod_{i=1}^{m} (f_i)^{d_i} \prod_{i=0}^{m} (S_i)^{c_i}. \tag{1.7}$$

Using equation (1.6), we can write equation (1.7) as

$$\begin{aligned}
L &= \prod_{i=1}^{m} \left[\lambda_i \prod_{j=1}^{i-1} (1-\lambda_j) \right]^{d_i} \prod_{i=1}^{m} \left[\prod_{j=1}^{i} (1-\lambda_j) \right]^{c_i}, \\
&= \prod_{i=1}^{m} \left[\left(\frac{\lambda_i}{1-\lambda_i} \right)^{d_i} \prod_{j=1}^{i} (1-\lambda_j)^{c_i+d_i} \right], \\
&= \prod_{i=1}^{m} \left(\frac{\lambda_i}{1-\lambda_i} \right)^{d_i} \prod_{i=1}^{m} \prod_{j=1}^{i} (1-\lambda_j)^{c_i+d_i}, \\
&= \prod_{i=1}^{m} \left(\frac{\lambda_i}{1-\lambda_i} \right)^{d_i} \prod_{i=1}^{m} \prod_{j=i}^{m} (1-\lambda_i)^{c_j+d_j},
\end{aligned}$$

and we finally obtain,

$$L = \prod_{i=1}^{m} \left(\frac{\lambda_i}{1-\lambda_i} \right)^{d_i} \prod_{i=1}^{m} (1-\lambda_i)^{\sum_{j=i}^{m} c_j + d_j}. \tag{1.8}$$

The log-likelihood is given by

$$l = \log L = \sum_{i}^{m} \left[d_i \log \lambda_i - d_i \log(1-\lambda_i) + \log(1-\lambda_i) \sum_{j=i}^{m} (c_j + d_j) \right].$$

So,

$$\frac{\partial l}{\partial \lambda_i} = \frac{d_i}{\lambda_i} + \frac{d_i}{1-\lambda_i} - \frac{\sum_{j=i}^{m} (c_i + d_i)}{1-\lambda_i}.$$

To obtain the maximum likelihood estimator, we equal the derivative to 0. Then, we obtain the solution that we desire:

$$\frac{d_i}{\widehat{\lambda}_i} + \frac{d_i}{1-\widehat{\lambda}_i} - \frac{\sum_{j=i}^{m} (c_i + d_i)}{1-\widehat{\lambda}_i} = 0 \Rightarrow \widehat{\lambda}_i = \frac{d_i}{\sum_{j=i}^{m} (c_j + d_j)}, \quad i = 1, ..., m.$$

□

Remark 1.2.2.1. The hazard at each t'_i is therefore estimated by the observed number of failures at t'_i as a proportion of the number at risk at t'_i.

1.2.3 Kaplan Meier estimator

From the discrete hazard maximum likelihood estimator, we can obtain an estimator for the survival function. Using equation (1.6), we get the *Kaplan-Meier estimator* of the survival function.

Definition 1.2.1. The *Kaplan-Meier* (or *product limit estimator*) estimator of the survival function is given by:

$$\widehat{S}(t) = \prod_{j=1}^{i} (1 - \widehat{\lambda}_j) \ , \ t \in [t'_i, t'_{i+1}).$$

The Kaplan-Meier estimator is the most widely used non-parametric estimator of the survival function. This method partitions the interval time and estimates the survival function on each separation. The derived estimates, $\hat{S}_i$, can then be used to construct the well-known Kaplan-Meier curve which provides a graphic illustration of the survival function.

The standard error of the Kaplan-Meier estimator can be obtained using *Greenwood's formula* in [5]

$$\sigma^2(t) = (\widehat{S}(t))^2 \sum_{j=1}^{i} \frac{d_j}{r_j(r_j - d_j)}, \quad t \in [t'_i, t'_{i+1}), i = 0, ..., m \tag{1.9}$$

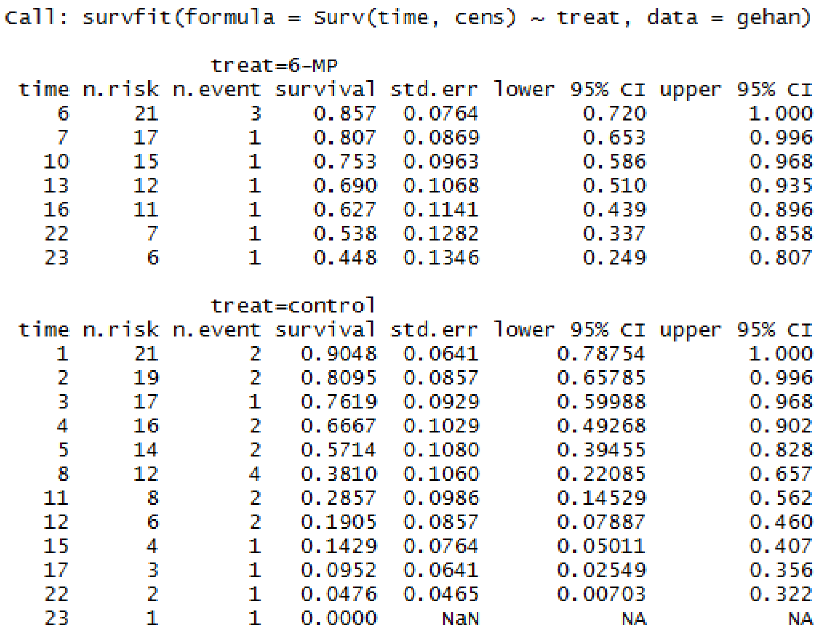
Example 1.2.3.1. To illustrate the Kaplan-Meier estimator, we will consider the Gehan data available in R using library(MASS). The Gehan data is a data frame of trials containing 42 leukaemia patients. Some of them were treated with the drug 6-mercaptopurine (6-MP) and the remainder received a placebo to become the control group. Table 1.2 below illustrates some of the Gehan data:

Table 1.2: Gehan data

patient	pair	time	cens	treatment
1	1	1	1	control
2	1	10	1	6-MP
3	2	22	1	control
4	2	7	1	6-MP
⋮	⋮	⋮	⋮	⋮
41	21	8	1	control
42	21	10	0	6-MP

The Kaplan-Meier estimates using 6-MP and the control group are presented in Figure 1.2 and the corresponding Kaplan-Meier estimate of $S(t)$ is depicted in Figure 1.3 below.

Figure 1.2 contains distinct time of observations ("time"), the number at risk ("n.risk"), the number of events ("n.events"), the standard error ("std.err") using *Greenwood's formula* and the confident interval in each group. Moreover,Figure 1.3 allows comparison of the two survival curves of each group. We can observe from the Kaplan-Meier curves that patients from the treatment group, in this example 6-MP, have a higher survival probability than patients from the control group.



```
Call: survfit(formula = Surv(time, cens) ~ treat, data = gehan)

                treat=6-MP 
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   6     21       3    0.857  0.0764        0.720        1.000
   7     17       1    0.807  0.0869        0.653        0.996
  10     15       1    0.753  0.0963        0.586        0.968
  13     12       1    0.690  0.1068        0.510        0.935
  16     11       1    0.627  0.1141        0.439        0.896
  22      7       1    0.538  0.1282        0.337        0.858
  23      6       1    0.448  0.1346        0.249        0.807

                treat=control 
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   1     21       2   0.9048  0.0641      0.78754        1.000
   2     19       2   0.8095  0.0857      0.65785        0.996
   3     17       1   0.7619  0.0929      0.59988        0.968
   4     16       2   0.6667  0.1029      0.49268        0.902
   5     14       2   0.5714  0.1080      0.39455        0.828
   8     12       4   0.3810  0.1060      0.22085        0.657
  11      8       2   0.2857  0.0986      0.14529        0.562
  12      6       2   0.1905  0.0857      0.07887        0.460
  15      4       1   0.1429  0.0764      0.05011        0.407
  17      3       1   0.0952  0.0641      0.02549        0.356
  22      2       1   0.0476  0.0465      0.00703        0.322
  23      1       1   0.0000     NaN           NA           NA
```

Figure 1.2: Output from R-code of the Kaplan-Meier estimates

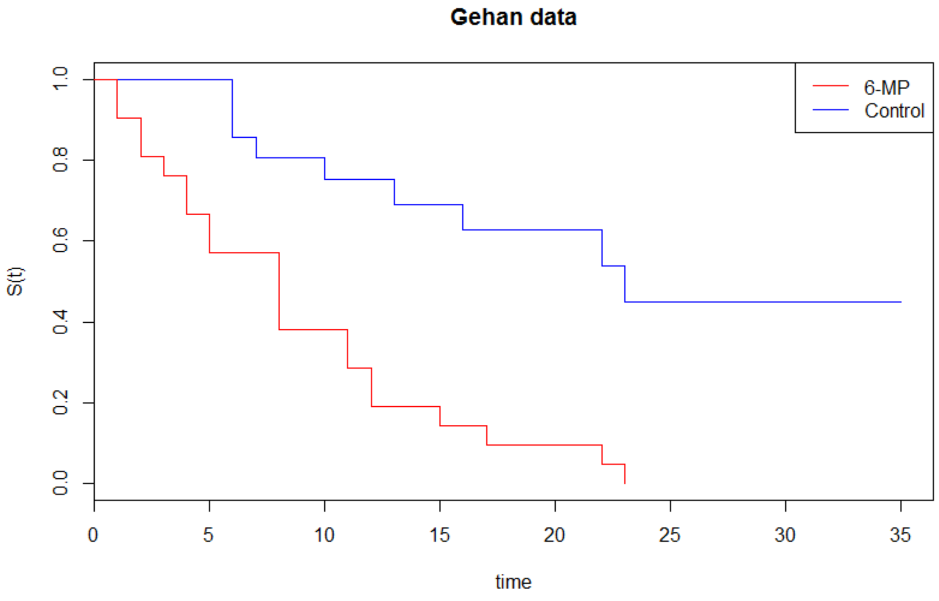


Figure 1.3: Kaplan-Meier curve using Gehan data

The comparison of survival plots allows us to define whether a treatment has an effect on the survival probability of patients or not. A first approach is to use the Kaplan-Meier plot of the different survival curves and look for differences at individual times. Although simple, this approach provides a preliminary idea of the comparison, but it is not enough to conclude that one group has a better survival rate than another group. A solution to test the difference between two or more survival curves is to use a statistical test.

Therefore, in Chapter 2, we describe different statistical tests which test the null hypothesis that there is no difference between the survival curves. We can write this null hypothesis and the alternative hypothesis as follows

$$\mathbf{H}_0 : S_{Trt}(t) = S_{Ctr}(t) \longleftrightarrow \mathbf{H}_1 : S_{Trt}(t) \neq S_{Ctr}(t), \tag{1.10}$$

where $\mathbf{H}_0$ is the null hypothesis and $\mathbf{H}_1$ is the alternative hypothesis.

The classic method testing the differences in survival curves is more powerful when the proportional hazard assumption holds. However, in the context of immunotherapy, the patterns of the survival curves violate the proportional assumption. In the final part of this chapter, we describe this proportional assumption.

1.3 Proportional assumption

The proportional assumption is defined as the fact that the ratio of the hazard function corresponding to both groups, treatment and control, is constant over time. This is illustrated in Figure 1.4. Moreover, the Figure 2.1, 2.2 and 2.3 are example where the proportional assumption is not checked.

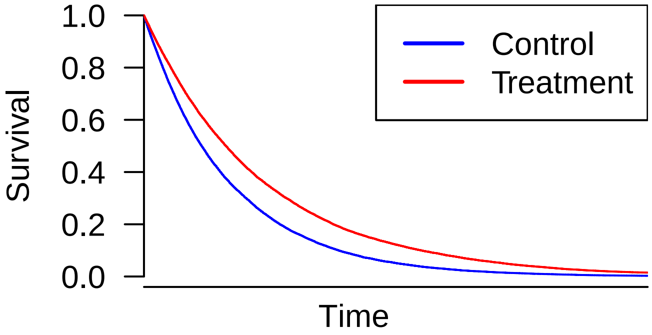


Figure 1.4: Survival plot illustrating proportional hazard

We can also define this assumption using the concept of *hazard ratio* (HR). Indeed, the *hazard ratio* is an estimate of the ratio between the hazard functions in the treatment and the control group [6]. If we think about the definition of hazard function 1.1.2, we can interpret the hazard ratio as the odds of an event occurring in the treatment group divided by the odds of the event appearing in the control group. The assumption of proportional hazard can be described by a hazard ratio constant over time.

The proportional hazards assumption may also be assessed informally by inspecting the Kaplan-Meier curves. However, this method may be misleading in some cases. Another possibility is the use of Grambsch and Therneau's approach to diagnose non-proportionality in [1]. Royston and Parmar used a likelihood ratio test to compare survival models with and without a time-dependent covariate. The proportional hazards assumption tests will not be discussed here as they go beyond the scope of this master's thesis.

Chapter 2

Survival methods

When the proportional hazard assumption is in doubt, the logrank test loses power. In order to solve this loss of power, some alternative methods have been developed. In this master's thesis, we will consider five alternative methods: *restricted mean survival time, generalised pairwise comparison, weighted logrank, adaptive weighted logrank* and *weighted Kaplan-Meier tests*. These methods have been selected due to their availability in R. However, many different methods such as the *cure model*[7] or *combined test*[8] exist as an alternative to the logrank.

In this chapter, we will define and present the logrank test and the statistical test of five methods introduced above. Moreover, we will give the R-function, which will be used in the next chapter and some properties of the different methods will be commented upon.

Some methods have different properties or can be adapted according to the non-proportional hazard trials. The three main non-proportional hazard situations include the early effect, the delayed effect and the reverse effect treatment over time [9]. It is possible to design other types of trials as a mixture of those three effects. Below, the based trials which do not respect the proportional hazard assumption are introduced.

The early effect
The early treatment effect is a treatment effect whose hazard ratio favours the treatment group. Indeed, as illustrated in Figure 2.4, the hazard ratio is smaller than 1 early in the trial and approaches or even exceeds 1 over time. We can observe this in the survival plot as the two curves are separated at the beginning of the study and remain so until the end. This is illustrated in Figure 2.1

The delayed effect
The late treatment effect is a treatment effect whose hazard ratio does not favour any group as HR=1 early in the trials and is inferior to 1 over time. This variation of the hazard ratio over time is illustrated in Figure 2.4. On the survival plot in Figure 2.2, the two curves are closed at the beginning of the study and they do separate, but later than

in the survival plot for the early effect. That remains so until the end of the study.

Reverse treatment effect over time
The reverse treatment effect over time occurs when the two survival curves cross over time. This is shown in Figure 2.3.

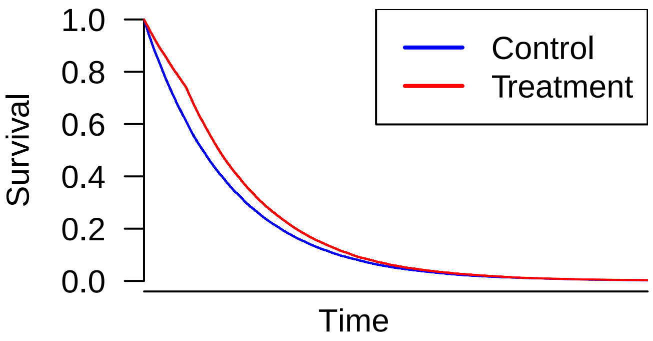


Figure 2.1: Survival plots illustrate the early effect

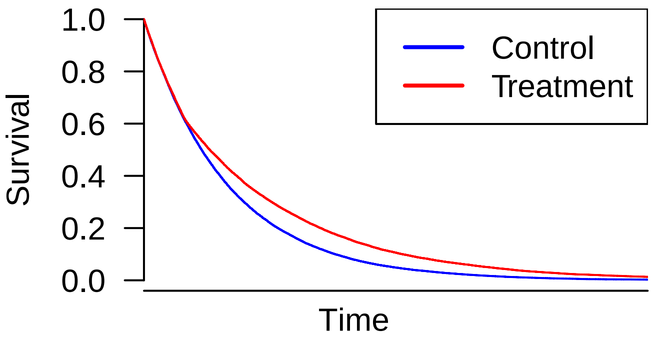


Figure 2.2: Survival plot illustrating the late effect

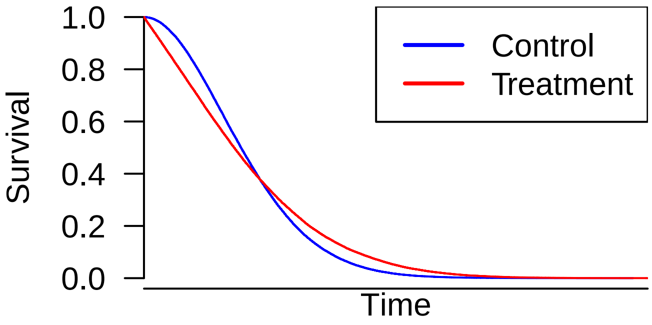


Figure 2.3: Survival plot illustrating the reverse treatment effect

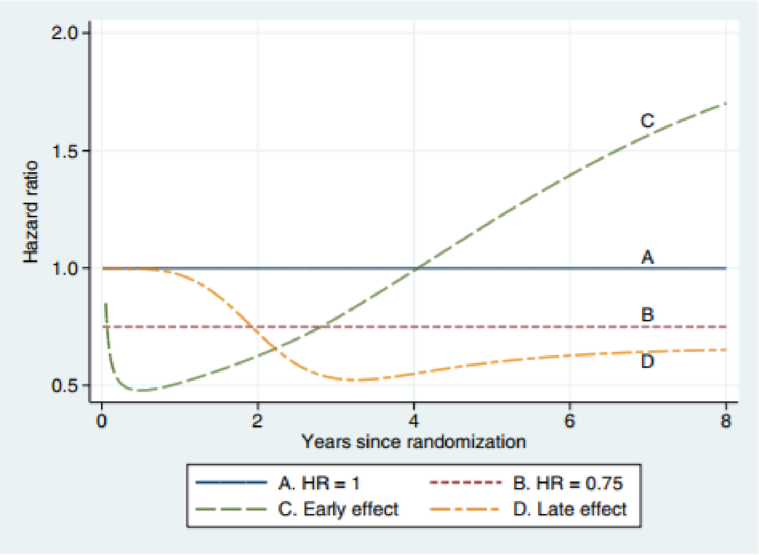


Figure 2.4: Hazard ratio according to different patterns

2.1 Logrank test

This section is written based on [10]. The logrank test is the most popular non-parametric test to compare two survival curves from two different groups. Compared to the Kaplan-Meier plot approach, this test takes the whole follow-up period into account. Therefore, it is not necessary to know anything about the shape of the survival curve or the distribution of the event.

In the following text, we will define the statistic for the logrank test and illustrate it with the example of the Gehan data.

Table 2.1 summarizes all the information from both groups, treatment (Trt) and control (Ctr), at a specific time.

Table 2.1: Summary of the information from both groups at a specific failure time

	Failed	Surviving	At risk
Treatment group	d_{Trt}	$n_{Trt} - d_{Trt}$	n_{Trt}
Control group	d_{Ctr}	$n_{Ctr} - d_{Ctr}$	n_{Ctr}
Combined	d	$n - d$	n

If the null hypothesis ($\mathbf{H}_0$) from relation 1.10 is true, i.e. the equality $S_{Trt}(t) = S_{Ctr}(t)$ is checked, then we expect that the total number of failures d is distributed between the two groups in the ratio of the group sizes; in other words, the expected number of failures in group Trt and group Ctr, respectively, at this time t_i :

$$e_{Trt_i} = \frac{n_{Trt_i} \times d_i}{n_i} \quad , e_{Ctr_i} = \frac{n_{Ctr_i} \times d_i}{n_i}.$$

We denote the total expected number of failures in group Trt and group Ctr respectively as follows:

$$E_{Trt} = \sum_i e_{Trt_i} \quad , E_{Ctr} = \sum_i e_{Ctr_i}.$$

To determine what happens at the next failure time $t_i + 1$, we need to update the contents of Table 2.1. The process of updating is done separately for each group and works as follows: the number at risk in the table at failure time t_{i+1} is the number at risk in the table at failure time t_i minus the number of failures in the table at failure time t_i minus the number of individuals censored after failure time t_i and before failure time t_{i+1}.

For the computation of the logrank statistic, we need the total of observed failures in each group:

$$O_{Trt} = \sum_i d_{Trt_i} \text{ and } O_{Ctr} = \sum_i d_{Ctr_i}.$$

The statistic for the logrank test is:

$$LR = X^2 = \frac{(O_{Trt} - E_{Trt})^2}{E_{Trt}} + \frac{(O_{Ctr} - E_{Ctr})^2}{E_{Ctr}}.$$

This statistic follows a chi-squared distribution with one degree of freedom. As a consequence, $\mathbf{H}_0$ is rejected if the observed value of X^2 is in the upper tail of the χ^2_1 distribution (where the degree of freedom is equal to the number of groups tested -1).

The logrank test is the most powerful non-parametric test under proportional hazard assumption and loses power when the curves cross.

This test is implemented in R using the command "survdiff" in the "survival" package.

Example 2.1.0.1. Let us continue with the Gehan example to illustrate the logrank test. First, we created Table 2.2, obtained by applying Table 2.1. Then, we calculated the logrank statistic which is given by

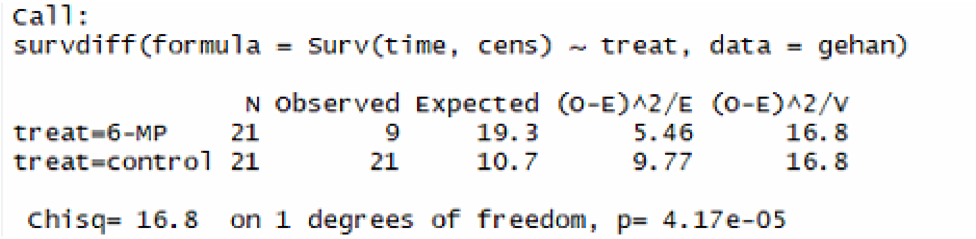
$$\begin{aligned} X^2 &= \frac{(-10.26)^2}{19.26} + \frac{(10.26)^2}{10.74} \\ &= 5.46 + 9.8 = 15.26. \end{aligned}$$

We compared this against a chi-squared distribution with 1 degree of freedom and we reject the null hypothesis at 5% level of significance because the $\chi^2_1(0.05) = 3.84$ and $X^2 > 3.84$.

Using the "survdiff" function in R represented on Figure 2.5, we get the same result. Indeed, if the p-value < 0.05 then we reject the null hypothesis. That means the survival functions are different. In this example, Figure2.5 shows that the p-value is such that $p < 0.0001$, so we clearly reject the null hypothesis.

Table 2.2: Logrank table for Gehan data

		# deaths		# in risk set		# expected		Observed-expected	
j	t'_j	d_{Trt_j}	d_{Ctr_j}	n_{Trt_j}	n_{Ctr_j}	e_{Trt_j}	e_{Ctr_j}	$d_{Trt_j} - e_{Trt_j}$	$d_{Ctr_j} - e_{Ctr_j}$
1	1	0	2	21	21	(21/42) × 2	(21/42) × 2	-1.00	1.00
2	2	0	2	21	19	(21/40) × 2	(19/40) × 2	-1.05	1.05
3	3	0	1	21	17	(21/38) × 1	(17/38) × 1	-0.55	0.55
4	4	0	2	21	16	(21/37) × 2	(16/37) × 2	-1.14	1.14
5	5	0	2	21	14	(21/35) × 2	(14/37) × 2	-1.20	1.20
6	6	3	0	21	12	(21/33) × 3	(12/33) × 3	1.09	-1.09
7	7	1	0	17	12	(17/29) × 1	(12/29) × 1	0.41	-0.41
8	8	0	4	16	12	(16/28) × 4	(12/28) × 4	-2.29	2.29
9	10	1	0	15	8	(15/23) × 1	(8/23) × 1	0.35	-0.35
10	11	0	2	13	8	(13/21) × 2	(8/21) × 2	-1.24	1.24
11	12	0	2	12	6	(12/18) × 2	(6/18) × 2	-1.33	1.33
12	13	1	0	12	4	(12/16) × 1	(4/16) × 1	0.25	-0.25
13	15	0	1	11	4	(11/15) × 1	(4/15) × 1	-0.73	0.73
14	16	1	0	11	3	(11/14) × 1	(3/14) × 1	0.21	-0.21
15	17	0	1	10	3	(10/13) × 1	(3/13) × 1	-0.77	0.77
16	22	1	1	7	2	(7/9) × 2	(2/9) × 2	-0.56	0.56
17	23	1	1	6	1	(6/7) × 2	(1/7) × 2	-0.71	0.71
Totals		9	21			19.26	10.74	-10.26	10.26



```
Call:
survdiff(formula = Surv(time, cens) ~ treat, data = gehan)

                N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP     21        9     19.3      5.46      16.8
treat=control  21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

Figure 2.5: Result of the logrank test using R

2.2 Restricted mean survival time

The *Absolute Difference Restricted Mean Survival time* is based on the difference of the area under the survival curves from time 0 to a pre-specified time point using normal approximation. This method is implemented in R in the survRM2-package using the "rmst2" function.

In this section, we define the *Restricted Mean Survival time* as given by Royston and Parmar [11].Then, the Kaplan-Meier estimate is presented to estimate the restricted mean survival time. Finally, the statistic test is developed, the R-function "rmst2" used for the simulation is detailed, and some properties of this method are presented.

Definition 2.2.1. The *Restricted Mean Survival Time* (RMST), denoted by μ_τ, of a random variable T, is the mean of survival time $X = min(T, \tau)$:

$$\mu_\tau = E[X] = \int_0^\tau S(t)\mathrm{d}t, \text{ where } \tau > 0. \tag{2.1}$$

When the survival time is years until death, we can interpret μ_τ as τ years of life expectancy. Moreover, the measure μ_τ increases monotonically with τ because the integral gives a non-negative, increasing function of τ. Indeed, $\int_0^{\tau_2} S(t)\mathrm{d}t > \int_0^{\tau_1} S(t)\mathrm{d}t$ when $\tau_2 > \tau_1$. In this state, nothing is known about the threshold τ.

The pre-specification of this threshold must be established before analysing the trials. This specification of τ is either defined by the context of the study or defined as the minimum of the largest observed time in each of the two groups.

Before defining the statistic test, we will introduce the Kaplan-Meier method of estimating the restricted mean survival time as described by Wei and Royston [12].
The most currently used method to estimate the restricted mean survival time is to directly integrate the Kaplan-Meier estimate of the survival function from time 0 to τ.

This integrate can be calculated using definition 1.2.1 by

$$\sum_{j=0,0\leq t_j\leqslant\tau}^{k} \hat{S}(t_j)(t_{j+1}-t_j),$$

where $\hat{S}(t_j)$ is the Kaplan-Meier estimate at time $t_j(0 \leqslant t_j \geqslant \tau)$ and t_j is the time at which an event occurs.

Other methods, such as pseudo-observations and flexible parameter models, of estimating the restricted mean survival time are described in appendix A.

In order to use the normal approximation for the statistic test, the restricted standard deviation of survival time is derived from the restricted mean survival time in [13].

The *Restricted Variance of Survival Time* is given by

$$var(X) = E[X^2] - (E[X])^2 = 2\int_0^\tau tS(t)\mathrm{d}t - \left(\int_0^\tau S(t)\mathrm{d}t\right)^2, \tag{2.2}$$

then *the Restricted Standard Deviation of Survival Time* is $\sqrt{var(X)}$.

To calculate the variance of the restricted survival time X, we need to compute $E[X^2]$. By the law of total probability, we get:

$$E[X^2] = E[T^2|T \leq \tau]\mathbb{P}(T \leq \tau) + \tau^2\mathbb{P}(T > \tau).$$

Moreover, we know by definition that $\mathbb{P}(T > \tau) = S(\tau)$ and $\mathbb{P}(T \leq \tau) = 1-S(\tau)$. Applying the definition of conditional expectation we get an expression for the first term of the sum:

$$\begin{aligned} E[T^2|T \leq \tau]\mathbb{P}(T \leqslant \tau) =& \int_0^\tau t^2 f(t)\mathrm{d}t, \\ =&\tau^2[1-S(\tau)] - \int_0^\tau 2t[1-S(t)]\mathrm{d}t. \end{aligned}$$

Hence, by replacing the first term by the one above, a shorter expression for $E[X^2]$ is derived

$$\begin{aligned} E[X^2] =&\tau^2[1-S(\tau)] - 2\int_0^\tau t[1-S(t)]\mathrm{d}t + \tau^2 S(\tau), \\ =&\tau^2 - 2\int_0^\tau t\mathrm{d}t + 2\int_0^\tau tS(t)\mathrm{d}t, \\ =&2\int_0^\tau tS(t)\mathrm{d}t. \end{aligned}$$

Finally, we get the equality (2.2) by subtracting the square of the value of the equality (2.1) from the last expression of $E[X^2]$.

To test the null hypothesis, i.e. there are no differences between survival curves, we define the difference in restricted mean survival time using the definition proposed by Lin and Xu [14].

Definition 2.2.2. The *difference in restricted mean survival time*(RMSTD) between two arms, denoted Δ, is given by

$$\Delta = \int_0^\tau S_{Trt}(t) - S_{Ctr}(t)\,\mathrm{d}t = \mu_{Trt} - \mu_{Ctr}.$$

where μ_{Trt} (resp. μ_{Ctr}) is the restricted mean survival time for the treatment group (resp. control group) with a threshold τ.

The interpretation of Δ is straightforward. Indeed, in the case of the time-scale being in years, the difference in RMST can be interpreted as patients in one group gaining or losing Δ more years in life expectancy from the origin to the threshold τ compared with patients in the other group. Moreover, a difference in restricted mean survival time greater than 0 favours the treatment group.

We can use another approach proposed by Zhao et al.[15] as an alternative measure to the hazard ratio. It is the relative difference in restricted mean survival time. It is given by the ratio between the RMSTD and τ.

$$\overline{\Delta} = \frac{\mu_{Trt} - \mu_{Ctr}}{\tau},$$

where μ_{Trt} (resp. μ_{Ctr}) is the restricted mean survival time for the treatment group (resp. control group) with a threshold τ. This measure quantifies how the RMSTD change with τ. Moreover, this quantity varies between 0 and 1 so it can be interpreted as a percentage.

We can use the Kaplan-Meier method to approximate the difference in restricted mean survival time as

$$\Delta = \sum_{j|t_j<\tau} \hat{S}_{Trt}(t_j) - \hat{S}_{Ctr}(t_j)(t_{j+1} - t_j). \tag{2.3}$$

We can reformulate the null hypothesis 1.10 as follow:

$$\mathbf{H}_0 : \mu_{Trt} = \mu_{Ctr} \longleftrightarrow \mathbf{H}_1 : \mu_{Trt} \neq \mu_{Ctr}. \tag{2.4}$$

Hence we have, $\Delta = 0$ under $\mathbf{H}_0$ and $\Delta \neq 0$ under $\mathbf{H}_1$. For the statistic test we estimate Δ and $var(\Delta)$ as follows:

$$\hat{\Delta} = \hat{\mu}_{Trt} - \hat{\mu}_{Ctr}$$

and

$$var(\hat{\Delta}) = \frac{\hat{\sigma}^2_{Trt}}{n_{Trt}} + \frac{\hat{\sigma}^2_{Ctr}}{n_{Ctr}}$$

where n_{Trt} and n_{Ctr} is the size of the treatment group and control group respectively. Moreover, $var(\hat{\mu}_{Trt})$ and $var(\hat{\mu}_{Ctr})$ are estimated using the delta method. This delta method is defined in appendix B.

The following test statistic is proposed by [13] for the comparison of the survival curves between two groups:

$$Z = \frac{\Delta}{\sqrt{var(\Delta)}}.$$

With the standardization of Δ, Z will be asymptotically standard normal. At significance level α,we reject the null hypothesis if the p-value is smaller than α where α is the level of the type I-error. Indeed, when $|Z| > Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is the quantile $1 - \alpha/2$ of the standard normal distribution, we reject the null hypothesis.

Now, we have the statistic test that is used in the R-function "rmst2", we can detail this function. This function uses six parameters: time, status, arm, tau, covariates and alpha. In this master thesis, we do not use covariates so the parameter covariates is initialised at NULL. In Chapter 3, we derive the type I-error at a level of significance $\alpha = 0.05$ then the value of the parameter alpha is 0.05. If we do not have a pre-specified threshold then the value of the parameter tau is defined as the minimum of the largest observed time in each of the two groups. The values time, status and arm correspond to the followed-up time for right censored data, the status indicator (event=1 and censored=0) and the group indicator for comparison (Trt=1 and Ctr=0) respectively.

Next, some properties established by Royston and Parmer[11] are introduced.

Properties 2.2.0.1.

- *Under proportional hazard, the RMSTD continues to increase with τ.*
- *Under proportional hazard, RMST gives a p-value comparable to that from the logrank test.*
- *The difference in RMST is a safer measurement because it is free of the proportional hazard assumption.*
- *The main advantages are the interpretability of the RMST difference from a clinical perspective such as loss of life expectancy, and the robustness of the estimator to the proportional hazard assumption.*
- *The main disadvantage is the dependence of the test statistic on τ.*

2.3 Generalised pairwise comparison

This section is written based on [16] and [17].

As the name indicates, the generalised pairwise comparison takes pairs of individuals from the two groups and compares them. In this section, we introduce some new notations in order to classify these pairs into four groups: "Favourable", "Unfavourable", "Neutral" and "Uninformative". Next, we introduce the net benefit which enables us to perform a test analysis of the null hypothesis. Then, we use an extensive procedure that takes into account the "Uninformative" pairs and updates the net benefit. Finally, we detail the R-function related to this method and introduce some properties.

Let n_{Trt} (resp. n_{Ctr}) denote the number of individuals who received treatment, denoted by Trt (resp. the number of individuals who are in the control group, denoted by Ctr). There is no prerequisite regarding groups formation, so they are formed by random allocation.

Pairwise comparisons consider pairs of individuals: one taken from the treatment group and the other one taken from the control group. Then the outcomes of these two individuals are compared and classified into one of the different categories: "Favourable, "Unfavourable", "Neutral", "Uninformative".

- "Favourable": If the outcome of the individual in the treatment group is better than the outcome of the individual in the control group.
- "Unfavourable": If the outcome of the individual in the treatment group is worse than the outcome of the individual in the control group.
- "Neutral": If there is no difference between the outcomes of the two individuals.
- "Uninformative": Otherwise (e.g. if the outcomes are censored).

In our case, as we are working with a time-to-event variable, we consider that the variable X (resp. Y) in the treatment group (resp. control group) can be right censored. Let ϵ and η denote the censoring indicator of variables X and Y respectively. When $\epsilon_i = 0$ (resp. $\eta_j = 0$) indicates that X_i (resp. Y_j) is a censored observation. This censored observation is denoted by X_i' and Y_j'. Let τ be the pre-specified threshold. As Ozenne et al. explain in [17], the threshold is used when "the difference between two variables needs to exceed a clinically relevent threshold". It can be a function of the precision with which X (resp. Y) is measured.

Table 2.3 shows how the different combinations of two outcomes can fill in the different categories.

Table 2.3: Generalized pairwise comparisons for a time-to-event

(ϵ_i, η_j)	Pairwise comparison	Pair is
(1,1)	$X_i - Y_j > \tau$	favourable
	$X_i - Y_j \leqslant -\tau$	unfavourable
	$\lvert X_i - Y_j \rvert < \tau$	neutral
(0,1)	$X_i' - Y_j > \tau$	favourable
	$X_i' - Y_j \leqslant -\tau$	uninformative
	$\lvert X_i' - Y_j \rvert < \tau$	uninformative
(1,0)	$X_i - Y_j' > \tau$	uninformative
	$X_i - Y_j' \leqslant -\tau$	unfavourable
	$\lvert X_i - Y_j' \rvert < \tau$	uninformative
(0,0)	$X_i' - Y_j' > \tau$	uninformative
	$X_i' - Y_j' \leqslant -\tau$	uninformative
	$\lvert X_i' - Y_j' \rvert < \tau$	uninformative

We define a pairwise indicator, for the pair formed by the i^{th} individual ($i = 1, ..., n_{Trt}$) in the treatment group and the j^{th} individual ($j = 1, ..., n_{Ctr}$) in the control group.

$$p_{ij} = \begin{cases} +1 & \text{if the pair is favourable} \\ 0 & \text{if the pair is neutral} \\ -1 & \text{if the pair is unfavourable.} \end{cases}$$

Now, thanks to the pairwise indicator, we can define the net difference which allows us to test the null hypothesis. The "proportion in favour of treatment", also called *net benefit*, denoted by $\Delta^\star$, is the net difference between the number of favourable pairs and the number of unfavourable pairs divided by the total number of pairs. Then, the net benefit can be expressed by

$$\Delta^\star = \frac{\sum_{i=1}^{n_{Trt}} \sum_{j=1}^{n_{Ctr}} p_{ij}}{n_{Trt} \cdot n_{Ctr}}.$$

We can interpret the net benefit according to its value as follows:

- If $\Delta^\star = 1$ then the treatment group is uniformly better than the control group.
- If $\Delta^\star = -1$ then the control group is uniformly better than the treatment group.
- If $\Delta^\star = 0$ then there is no net difference between the groups.

As the last interpretation indicates, if there are no net differences between groups, then $\Delta^\star = 0$. Hence, we can express the null hypothesis that there are no differences between survival curves as the net benefit equals 0.

To test this new null hypothesis, a randomization test can be used. To do so, we need a large number, say S, of identical simulations to the real experiment under analysis. Hence, all individuals' data remains unchanged except the allocation of their treatment group (treatment or control), which is randomly re-allocated. Therefore, we define $\Delta_i^\star$ which is the proportion in favor of treatment in the i^{th} simulated experiment.

The type I-error of the randomization test can be calculated on the asymptotic normality of the empirical distribution on $\Delta^\star$ under the new null hypothesis i.e. $\Delta^\star = 0$. The standard deviation of the empirical distribution of $\Delta^\star$ is

$$\sigma = \frac{1}{2}\frac{(\Delta^\star_{1-\frac{\alpha}{2}} - \Delta^\star_{\frac{\alpha}{2}})}{Z_{1-\alpha}}$$

where $Z_{1-\alpha}$ is the $(1-\alpha)$ percent standardized normal variable, $\Delta^\star_{\frac{\alpha}{2}}$ is the value of $\Delta_i^\star(i = 1, ..., S)$ that leaves at most $\alpha/2$ percent values of $\Delta_i^\star$ to its left (i.e. $\Delta_i^\star \leq \Delta^\star_{\frac{\alpha}{2}}$) and $\Delta^\star_{1-\frac{\alpha}{2}}$ the value that leaves at most $\alpha/2$ percent values of $\Delta_i^\star$ to its right (i.e. $\Delta^\star_{1-\frac{\alpha}{2}} \leq \Delta_i^\star$).

Then the p-value is equal to $\Phi(-\frac{\Delta^\star_{obs}}{\sigma})$ for a one-sided test and 2 $\Phi(-\frac{\Delta^\star_{obs}}{\sigma})$ for a two-sided test, where $\Phi(.)$ is the standard normal cumulative distribution function.

All we have done to date does not take into account the "Uninformative" pairs which often occur when both outcomes are censored. Therefore, the net benefit can be reviewed to include the "Uninformative" pairs in the test. Two new notations are introduced, time-to-event denoted by x_i^0 and y_j^0 and time-to-observation denoted by x_i and y_j. We can update the definition of the censoring indicator as follows:

$$\epsilon_i = \begin{cases} 1 & \text{if } x_i = x_i^0 \\ 0 & \text{if } x_i < x_i^0 \end{cases}$$

$$\eta_i = \begin{cases} 1 & \text{if } y_j = y_j^0 \\ 0 & \text{if } y_j < y_j^0. \end{cases}$$

Then we can also update the definition of survival function with the new notation as

$$S_{Trt}(t) = \mathbb{P}[x_i^0 \geqslant t] \text{ and } S_{Ctr}(t) = \mathbb{P}[y_j^0 \geqslant t].$$

Based on the Kaplan-Meier estimate of the survival function, we get

$$\mathbb{P}[x_i^0 > t | x_i, \epsilon_i = 0] = \frac{\hat{S}_{Trt}(t)}{\hat{S}_{Trt}(x_i)}.$$

Then, we calculate a new pairwise score for each combination of $\{x_i, y_j, \epsilon_i, \eta_j\}$:

$$s_{ij} = \mathbb{P}(x_i^0 > y_j^0 + \tau) - \mathbb{P}(y_j^0 > x_i^0 + \tau).$$

When a pair can be categorised as favourable or unfavourable then s_{ij} takes the value -1 or 1 and for other pairs, $s_{ij} \in [-1, 1]$. We can then build Table 2.4 which shows all the values of the pairwise score according to each combination of $\{x_i, y_j, \epsilon_i, \eta_j\}$. In appendix C, all the details concerning the formulae in each row are explained.

Table 2.4: Value of s_{ij} for pairwise comparison of a time-to-event outcome.

(ϵ_i, η_j)	$x_i - y_j > \tau$	$x_i - y_j < -\tau$	$\lvert x_i - y_j \rvert < \tau$
$(1,1)$	1	-1	0
$(0,1)$	1	$\frac{\hat{S}_{Trt}(y_j+\tau)+\hat{S}_{Trt}(y_j-\tau)}{\hat{S}_{Trt}(x_i)} - 1$	$\frac{\hat{S}_{Trt}(y_j+\tau)}{\hat{S}_{Trt}(x_i)}$
$(1,0)$	$1 - \frac{\hat{S}_{Ctr}(x_i+\tau)+\hat{S}_{Ctr}(x_i-\tau)}{\hat{S}_{Ctr}(y_i)}$	-1	$-\frac{\hat{S}_{Ctr}(x_i+\tau)}{\hat{S}_{Ctr}(y_i)}$
$(0,0)$	$1 - \frac{\hat{S}_{Ctr}(x_i-\tau)}{\hat{S}_{Ctr}(y_j)}$ $-\int_{\mathcal{B}}^{\infty} \frac{\hat{S}_{Trt}(t+\tau)}{\hat{S}_{Trt}(x_i)\hat{S}_{Ctr}(y_j)} \mathrm{d}\hat{S}_{Ctr}(t)$ $+\int_{\mathcal{C}}^{\infty} \frac{\hat{S}_{Ctr}(t+\tau)}{\hat{S}_{Trt}(x_i)\hat{S}_{Ctr}(y_j)} \mathrm{d}\hat{S}_{Trt}(t)$	$-\int_{\mathcal{A}}^{\infty} \frac{\hat{S}_{Trt}(t+\tau)}{\hat{S}_{Trt}(x_i)\hat{S}_{Ctr}(y_j)} \mathrm{d}\hat{S}_{Ctr}(t)$ $-\int_{\mathcal{D}}^{\infty} \frac{\hat{S}_{Ctr}(t+\tau)}{\hat{S}_{Trt}(x_i)\hat{S}_{Ctr}(y_j)} \mathrm{d}\hat{S}_{Trt}(t)$ $+\frac{\hat{S}_{Trt}(y_j-\tau)}{\hat{S}_{Trt}(x_i)} - 1$	$-\int_{\mathcal{A}}^{\infty} \frac{\hat{S}_{Trt}(t+\tau)}{\hat{S}_{Trt}(x_i)\hat{S}_{Ctr}(y_j)} \mathrm{d}\hat{S}_{Ctr}(t)$ $+\int_{\mathcal{C}}^{\infty} \frac{\hat{S}_{Ctr}(t+\tau)}{\hat{S}_{Trt}(x_i)\hat{S}_{Ctr}(y_j)} \mathrm{d}\hat{S}_{Trt}(t)$

where $\mathcal{A} = \{t > y_j : t \in \{y_j\} \text{ and } \delta_j = 1\}$, $\mathcal{B} = \{t > x_i - \tau : t \in \{y_j\} \text{ and } \delta_j = 1\}$, $\mathcal{C} = \{t > x_i : t \in \{x_i\} \text{ and } \delta_i = 1\}$ and $\mathcal{D} = \{t > y_j - \tau : t \in \{x_i\} \text{ and } \delta_j = 1\}$.

This score can be seen as the probability of a random patient in the treatment group having a "better outcome" than a random patient in the control group minus the opposite probability. The updated new net benefit is given by

$$\Delta^\star = \frac{1}{n_{Trt} \cdot n_{Ctr}} \sum_{i=1}^{n_{Trt}} \sum_{j=1}^{n_{Ctr}} s_{ij}.$$

Then, we apply the same method to calculate the p-value by replacing $\Delta^\star$ by the updated one.

In the next chapter, in the simulation study, we use the R-function "BuyseTest" with the following parameters: data, treatment, endpoint, type, threshold, censoring, method.tte, n.resampling, cpus. The parameter treatment, endpoint and censoring correspond to the name of the column corresponding respectively to treatment group, time and event. As already mentioned, if we work with time-to-event observation, then the parameter type is equal to "timeToEvent". We choose two thresholds, the first one is equal to 0 and the second to 1. We take the method of "Peron" for the parameter method.tte because it deals with uninformative pairs. We use one thousand resampling then the parameter n.resampling equals to 1000. We use the maximum of cpus, so the parameter cpus is initialised at "all". This method is denoted according to the value of the threshold: if the threshold is 0 (resp. 1), we denoted it by BT0 (resp. BT1).

During the simulation, we discovered some types of data where the R-function "BuyseTest" produced some errors. Discussion with L. Peron and M. Ozenne resulted is then discovering an inconsistency in the code revealed by our data and they improved the R-function correspondingly.

Below, some properties derived by Peron, Buyse et al. are introduced in [17].

Properties 2.3.0.1.

- *Under proportional hazard, the extended method is less powerful than the logrank test.*
- *This method depends on the threshold which must be predefined according to the clinical context.*
- *In the case of non-proportional hazards, during some simulations, authors of [17] have noticed that when the treatment effect increases over time, the extended procedure is more powerful than standard procedure. However, when the treatment effect decreases over time, the extended procedure is less powerful than other standard procedures.*
- *The extended procedure is more efficient to detect time-to-event difference than the standard procedure of generalised pairwise comparison when the proportional hazard assumption holds or in case of late effect.*

2.4 Weighted logrank test

This section is written based on Peckova and Fleming [18] and Lee [19].

As the logrank test is the most powerful test under proportional hazard assumption, the weighted logrank test can be chosen when the proportional hazard assumption is in doubt. In such situations, the weight can be selected in order to maximize the power.

In this section, we first introduce some notations. Then we define the statistic test and the different types of weight function that we can apply according to the shape of the curve, i.e. early effect or late effect. Finally, we detail the R-function that we used in the simulation and some properties are presented.

Consider T_{ij} the independent, positive random variables where i and j are the subscripts that respectively relate to the groups (i.e. $i \in \{Trt, Ctr\}$) and a specific time-to-event according to the groups (i.e. $j = 1, ..., n_i$). Let C_{ij} be the independent censoring variables which are also independent of the survival time T_{ij}. The available data consist of the pair (X_{ij}, δ_{ij}), $i \in \{Trt, Ctr\}$, $j = 1, ..., n_i$ where $X_{ij} = min(T_{ij}, C_{ij})$, and $\delta_{ij} = \mathbb{I}(T_{ij} \leqslant C_{ij})$ is the censoring indicator.

Let denote $\overline{N_i}$ be the number of failures in group i before or at time t and $\overline{Y_i}$ be the number at risk in group i at time t. These two values are defined as follows:

$$N_{ij}(t) = \delta_{ij}\,\mathbb{I}(X_{ij} \leqslant t), \quad \overline{N_i}(t) = \sum_{j=1}^{n_i} N_{ij}(t)$$

$$Y_{ij}(t) = \mathbb{I}(X_{ij} \geqslant t), \quad \overline{Y_i}(t) = \sum_{j=1}^{n_i} Y_{ij}(t).$$

We can correspond $\overline{N_i}$ to O_i and $\overline{Y_i}$ to the sum of n_i over time when we use the logrank notations.

Now we have all the notations to define the weighted logrank test statistic. Therefore, the *weighted logrank statistic* can be expressed for all $t \geqslant 0$ as

$$WLR = \sqrt{\frac{n_{Trt} + n_{Ctr}}{n_{Trt}\, n_{Ctr}}} \int_0^\infty W(t) \frac{\overline{Y}_{Trt}(t)\overline{Y}_{Ctr}(t)}{\overline{Y}_{Trt}(t) + \overline{Y}_{Ctr}(t)} \left(\frac{\mathrm{d}\overline{N_{Trt}}(t)}{\overline{Y_{Trt}}(t)} - \frac{\mathrm{d}\overline{N_{Ctr}}(t)}{\overline{Y_{Ctr}}(t)} \right), \tag{2.5}$$

where W(t) is a bounded nonnegative weight function.

The estimator of the variance of W under the null hypothesis, i.e. there is no difference between survival curves, is given by [18] and [20]:

$$\hat{\sigma}^2 = \int_0^\infty K^2(t) \frac{\overline{Y}_{Trt}(t) + \overline{Y}_{Ctr}(t)}{\overline{Y}_{Trt}(t)\overline{Y}_{Ctr}(t)} \left(1 - \frac{\Delta\overline{N_{Trt}}(t) + \Delta\overline{N_{Ctr}}(t) - 1}{\overline{Y}_{Trt}(t) + \overline{Y}_{Ctr}(t) - 1}\right) \times \frac{\mathrm{d}(\overline{N_{Trt}}(t) + \overline{N_{Ctr}}(t))}{\overline{Y}_{Trt}(t) + \overline{Y}_{Ctr}(t)} \tag{2.6}$$

where $K(t) = \sqrt{\frac{n_{Trt} + n_{Ctr}}{n_{Trt} n_{Ctr}}} W(t) \frac{\overline{Y}_{Trt}(t)\overline{Y}_{Ctr}(t)}{\overline{Y}_{Trt}(t) + \overline{Y}_{Ctr}(t)}$ and $\Delta\overline{N_i}(t) = \overline{N_i}(t) - \lim_{u \to t} \overline{N_i}(u)$.

Then, the following test statistic is proposed by [18] for the comparison of two survival curves:

$$X^2 = \frac{(WLR)^2}{\hat{\sigma}^2}. \tag{2.7}$$

At significance level α, we reject the null hypothesis when $X^2 > \chi^2_{1-\alpha}$ where $\chi^2_{1-\alpha}$ is the quantile $1 - \alpha$ of the Chi-squared distribution.

In the chapter 3, the Fleming and Harrington's weight [21] is used as $W(t)$ in equation (2.5) for the simulation study. This weight can be adaptive according to the shape of the curves. We obtain the statistic test proposed by Lee [19] when we replace $W(t)$ in the equation (2.5) by the Fleming-Harrington weight

$$W(t) = (\hat{S}(t-))^{\rho}(1 - \hat{S}(t-))^{\gamma}, \tag{2.8}$$

with $\hat{S}(t-)$ being the left-continuous Kaplan-Meier estimate for the survival function based on the pooled survival data [22], the exponents, ρ and γ, can take different values according to whether we want to put more weight on early departure or late departure. Indeed, when $\rho = 0$ and $\gamma = 0$ we obtain the logrank statistic test exactly. If we take $\rho = 1$ and $\gamma = 0$ then it gives more weight to early departure, whereas if we take $\rho = 0$ and $\gamma = 1$ then it gives more weight to departure which occurs later in time.

In the next chapter, we use the R-function "wtdlogrank" with the following parameters: formula, data, WtFun, param, sided. The formula corresponds to $Surv(time, event) \sim trt$ where the function "Surv" creates a survival object according to the treatment. The parameter data correspond to the dataset that we use for the simulation. The parameter WtFun equals "FH" is reference to Fleming Harrington's weight. Then the parameter param is related to the parameter WtFun. Indeed, the Fleming Harrington's weight uses two parameters: ρ and γ. So the parameter param specifies the value of ρ and γ. In the simulation study, we use $(\rho, \gamma) = (0, 1)$ and we denote this method by G01. Furthermore, we use $(\rho, \gamma) = (1, 0)$ and we denote this method by G10. Finally, the value of the parameter sided is 2.

Below, some properties based on simulation studies and example in [19] are introduced.

Properties 2.4.0.1.

- *Under proportional hazard, the logrank test, G01 and G10 achieve the nominal* 0.05 *level in the configuration with different sample size (20, 50 and 70) and with various censoring patterns (*0%, 20%, 40% *and* 60%*).*

- *Authors of [19] confirm that the logrank test, written using* $W(t)$ *with* $\rho = 0$ *and* $\gamma = 0$*, is the most powerful test when the proportional hazard is checked.*

- *When the configuration illustrates an early effect, G10 is the most powerful test in any case of sample and censoring patterns. Authors of [19] have observed the same with G01 in the configuration, illustrating a late effect.*

- *On a real application, the survival curves cross and authors of [19] have obtained results indicating that G10 does not reject the null hypothesis* (1.10) *whereas G01 shows a significant difference in the two groups.*

- *Detecting which value of* ρ *and* γ *to be used is still an issue.*

2.5 Adaptive weighted logrank test

This section is written based on Yang and Prentice [23].

When the hazard ratio is not proportional, the adaptive weight gives back the variance from proportionality and produces an improvement in power compared to the logrank test. The adaptive weights are obtained by fitting the model of Yang and Prentice to the data.

In this section, the same notations as above are used, and the model developed by Yang and Prentice, used to define the adaptive weight, is described. Then, a statistical test to evaluate the null hypothesis is proposed and the R-function used in the next chapter is described. Finally, some properties are introduced.

The model of Yang and Prentice [24] enables us to provide a more accurate description of the data in certain non-proportional hazard situations. This model can be developed as follows:

Let λ_{Ctr} and λ_{Trt} denote the hazard function for the two groups respectively and suppose that these functions belong to a parametric family $\{\lambda_\theta, \theta \in \Theta\}$. Yang and Prentice proposed a model in which

$$\lambda_{Trt}(t) = \frac{\theta_1\theta_2}{\theta_1 + (\theta_2 - \theta_1)S_{Ctr}(t)}\lambda_{Ctr}(t), \quad t < \tau_0 = \sup\{t : S_{Ctr}(t) > 0\} \tag{2.9}$$

where θ_1 and θ_2 are positive.

Under this new method, the hazard ratio between the two groups is non-constant. Indeed, at time $t < \tau_0$ it is given by

$$\frac{\lambda_{Trt}(t)}{\lambda_{Ctr}(t)} = \frac{\theta_1\theta_2}{\theta_1 + (\theta_2 - \theta_1)S_{Ctr}(t)}, \tag{2.10}$$

which clearly depends on θ_1, θ_2 and $S_{Ctr}(t)$. Notice that if $\theta_1 = \theta_2$ then we get a constant ratio that leads to a proportional hazard model. Moreover, this ratio is monotonically increasing if $\theta_2 > \theta_1$ and inversely, monotonically decreasing in the inverse case, i.e. $\theta_1 > \theta_2$. Under this model,

$$\theta_1 = \lim_{t\downarrow 0}\frac{\lambda_{Trt}(t)}{\lambda_{Ctr}(t)} \text{ and } \theta_2 = \lim_{t\uparrow\tau_0}\frac{\lambda_{Trt}(t)}{\lambda_{Ctr}(t)}.$$

We can interpret θ_1 and θ_2 as the short term and long-term hazard ratio respectively. When $\theta_1 \neq \theta_2$ and one is included in the interval formed by θ_1 and θ_2, the two hazard functions cross.

To test the hypothesis of no-treatment effect, we can define $\beta_1 = log(\theta_1)$ and $\beta_2 = log(\theta_2)$, and take $\beta_1 = \gamma_1\theta$ and $\beta_2 = \gamma_2\theta$. The test is the form of equality (2.5) with weight function

$$\Omega = \gamma_1 S_{Ctr} + \gamma_2(1 - S_{Ctr}) \tag{2.11}$$

We notice that we get the same result for the proportional hazard model if we take $\gamma_1 = \gamma_2$; this implies that Ω is a constant, in which case the logrank test is optimal.

In Appendix D, we show that

$$\Omega = \lim_{\theta \to 0} \frac{(1 - \frac{\lambda_{Ctr}}{\lambda_{Trt}})}{\theta} \tag{2.12}$$

then we can take as a weight:

$$\Omega_1 = \frac{\widehat{\lambda_{Ctr}}}{\widehat{\lambda_{Trt}}}$$

where the estimator of the hazard functions is calculated by fitting the model of Yang and Prentice to the data. Hence, this weight depends on the estimated hazard ratio. We can also define a new weight: $\Omega_2 = 1/\Omega_1$. The technical details to calculate the weight function are given in [23].

From [23], we can use the following test to reject the null hypothesis $\mathbf{H}_0$ when

$$\max(|X^2_{\Omega_1}|, |X^2_{\Omega_2}|) > Z_{1-\alpha/2}$$

where $X^2_{\Omega_1}$ and $X^2_{\Omega_2}$ is the standardised statistic from equation (2.7) with weights Ω_1 and Ω_2 respectively, and $Z_{1-\alpha/2}$ is the upper $100(1-\alpha/2)^{th}$ percentile of the standard normal distribution.

The main property described in [23] is that the adaptive method has better power than the logrank test under non-proportional hazard alternatives.

In the next chapter, we use the R-function "YPmodel.adlgrk" using only the dataset. This method is denoted by the abbreviation AWLR.

Below, we show properties which result from simulation studies and example in [23]

Properties 2.5.0.1.

- *This method is more powerful than the logrank test across several non-proportional hazard configurations such as early and late effects, and cross survival curves.*
- *The main advantage of this method is that we obtain the logrank test if the hazard ratio is constant, i.e. if the proportional hazard is checked.*

2.6 Weighted Kaplan-Meier test

Pepe and Fleming introduced the weighted Kaplan-Meier test in 1989-1991[25]. In this section, based on Uno et al. [26], we will describe the updating of the weighted Kaplan-Meier test. First, we will define Pepe and Fleming's statistic test then we will derive from this test two other statistic tests. Finally, we describe the R-function used in the next chapter and introduce some properties.

The weighted Kaplan-Meier statistic test developed by Pepe and Fleming is

$$WKM = \sqrt{\frac{n_{Ctr}\, n_{Trt}}{n_{Ctr} + n_{Trt}}} \int_0^{\tau} W(t)\hat{D}(t)\mathrm{d}t, \tag{2.13}$$

where $\hat{D}(t) = \hat{S}_{Trt}(t) - \hat{S}_{Ctr}(t)$, $\hat{S}_{Trt}(\cdot)$ and $\hat{S}_{Ctr}(\cdot)$ are the Kaplan-Meier estimators for the treatment and control group respectively. $\tau = \sup \Big[t : \min\{\hat{K}_{Ctr}(t), \hat{K}_{Trt}(t)\} > 0\Big]$[1], $\hat{K}_i(.)$ is the left-continuous version of the Kaplan-Meier estimator for the censoring survival function for the i^{th} groups, n_i is the sample size in group i with $i \in \{Trt, Ctr\}$ and $W(.)$ is the data-dependent weight function. Pepe and Fleming considered two weighting designs:

$$W_1(t) = \frac{\hat{K}_{Ctr}(t)\hat{K}_{Trt}(t)}{\hat{q}_{Ctr}\hat{K}_{Ctr}(t) + \hat{q}_{Trt}\hat{K}_{Trt}(t)}, \tag{2.14}$$

and

$$W_2(t) = \sqrt{\frac{\hat{K}_{Ctr}(t)\hat{K}_{Trt}(t)}{\hat{q}_{Ctr}\hat{K}_{Ctr}(t) + \hat{q}_{Trt}\hat{K}_{Trt}(t)}}, \tag{2.15}$$

where $\hat{q}_i$ is the proportion of subjects belonging to group i. This weighting scheme depends only on the censoring distribution.

Taking $W(t) = \hat{D}(t)$ in this way, we put more weight at time t where the difference between the two curves is "large". However, the distribution of the statistic test is similar to a "Chi-square" statistic and tends to have a rather long right tail under the null hypothesis, i.e. there is no difference between two survival curves.

In the following, the aim is to obtain a statistic test distribution such that, under the null hypothesis, this distribution has a short tail and under other alternatives, the observed test statistic tends to be large and then rejects the null hypothesis.

We consider again, as in section 2.4, the available data as the pair (X_{ij}, δ_{ij}), $i \in \{Trt, Ctr\}, j = 1, ..., n_i$, where $X_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = \mathbb{I}(T_{ij} \leqslant C_{ij})$ is the censoring indicator. Let $D(t) = S_{Trt}(t) - S_{Ctr}(t)$ denote the difference between curves at any time and $[0, \varsigma]$ be a given time interval. Furthermore, we consider that $\mathbb{P}(X_i > \varsigma) > 0, i \in \{Trt, Ctr\}$. The null hypothesis can be written as $D(t) = 0$, for $t \in [0, \varsigma]$. Now, let $\hat{D}(\cdot) = \hat{S}_{Trt}(\cdot) - \hat{S}_{Ctr}(\cdot)$, $\hat{\sigma}(\cdot)$ be its standard error estimate. Then $Z(\cdot) = \hat{D}(\cdot)/\hat{\sigma}(\cdot)$ has an asymptotic standard normal distribution under the null hypothesis. We replace $\hat{D}(t)$ in the equation (2.13) by $Z(t)$, then we obtain a new statistic test

$$V = \int_0^{\varsigma} W(t)Z(t)\mathrm{d}t, \tag{2.16}$$

[1] We can consider τ as the minimum of the largest censored observation in each of the two groups.

where $W(\cdot)$ is again a data-dependent weight function. Notice that we define $Z(t) = 0$ for $\hat{D}(t) = \hat{\sigma}(t) = 0$ that means $\hat{S}_{Trt}(t) = \hat{S}_{Ctr}(t) = 1$. That happens when no events occur at time t. With this statistic test, we still have a distribution similar to a "chi-square" with a rather long right tail under the null hypothesis and which for specific alternatives does not work well. When $W(.)$ is a constant function, the distribution of the test is more a standard normal distribution with its center around zero and a short tail under the null hypothesis. However, this is powerful when $Z(t)$ is constant over $[0, \varsigma]$. In this state, we need to find a weight function such that the statistic test distribution under the null hypothesis has a short tail but the observed V is larger under the alternative. Uno et al. suggested two solutions to respect the request of the statistic test distribution.

First, let $c \in [0, \eta]$, where η is a constant[2]. Let $W_c(t) = \max\{Z(t), c\}$ and

$$V_1(c) = \int_0^{\varsigma} W_c(t) Z(t) \mathrm{d}t. \tag{2.17}$$

In practice, c is usually fixed at 1.65. The choice of c is not obvious *a priori*; indeed the choice of $c = 1.65$ as made above may not work well in a case where $D(.)$ is positive for a large portion of time points then $Z(.)$'s are less than 1.65.

In appendix E, we show that fix c at 1.65 can be good choice in certain case.
The choice of c is not obvious *a priori*; indeed the choice of $c = 1.65$ as made above may not work well in a case where $D(.)$ is positive for a large portion of time points then $Z(.)$'s are less than 1.65.

Therefore, Uno et al. in [26] propose an approach to choose c adaptively in order to build a new test statistic based on $\{V_1(c), 0 \leqslant c \leqslant \eta\}$.

In appendix E, this approach of the choice of c is explained in general way. In practice, we take $\{\xi_{ij}, i \in \{Trt, Ctr\}, j = 1, ..., n_i\}$ a random sample from a standard normal distribution and the distribution of $V_1(c)$ under the null hypothesis can be approximated by generating M sets of $\{\xi_{ij}\}$. Then for each realised set $\{\xi_{ij}\}$, we calculate

$$V_1^{\star}(c) = \int_0^{\varsigma} W_c^{\star}(t) Z^{\star}(t) \mathrm{d}t, \tag{2.18}$$

where $Z^{\star}(\cdot) = (Q_T(\cdot) - Q_C(\cdot))/\hat{\sigma}(\cdot)$ and $W_c^{\star}(\cdot) = \max\{Z^{\star}(\cdot), c\}$.

We denote $\mathcal{D}$ as the set of M realisations of $\{V_1^{\star}(c), c \in [0, \eta]\}$ which is considered as a reference set for the new test. Then, using the reference set $\mathcal{D}$, we get the corresponding $P(c)$ [3] denoted by $P^{\star}(c)$. Furthermore, the distribution of $P_b = \min\{P(c) : c \in [0, \eta]\}$ under the null hypothesis can be estimated using the M realisations of $P_b^{\star} = \min\{P^{\star}(c) :$

[2] η is a positive constant, usually $\eta = 4$ works well.

[3] Where $P(c) = S_{V_1(c)}(V_1(c))$ and $S_{V_1(c)}(v)$ is the survival function of $V_1(c)$

$0 \leqslant c \leqslant \eta\}$ which is based on the set $\{V_1^\star(c), c \in [0, \eta]\}$.

The bona fide p-value of the new test is then given by $\mathbb{P}(P_b^\star < p_b)$ where $p_b = \min\{P(c) : c \in [0, \eta]\}$.

Another statistic test proposed by Uno et al. in [26] is

$$V_2(c) = \int_0^\varsigma W_c(t) Z(t) \mathrm{d}\overline{N}(t), \tag{2.19}$$

where $\overline{N}(t) = \dfrac{\sum_{i=T}^{C} \sum_{j=1}^{n_i} \mathbb{I}(X_{ij} \leqslant t)\delta_{ij}}{n_{Trt} + n_{Ctr}}$.

$\overline{N}(t)$ in the equation (2.19) is used as a weighting function of $Z(\cdot)$. Indeed, the weight is huge when there are many observed events in the time intervals. We can use the same approach to obtain the bona fide p-value corresponding to $V_2(c)$.

In the next chapter, we use the R-function "survAWKMT2" with the following parameters: indata, tau, nmethod, seed, v1, v2, test. The parameter indata corresponds to the dataset. The parameter tau equals the maximum of the observed time. nmethod corresponds to the number of resampling and for the simulation study, 1000 iterations for the resampling are choosen. We take the value b*123+123 for the seed where b corresponds to b^{th} iteration in the simulation study. v1 and v2 are initialised at TRUE and test at "2-side". Then we obtain the p-value of each statistic test. We denote V1 and V2 the abbreviation corresponding to the statistic test in the equations (2.17) and (2.19). Moreover, we denote BFV1 and BFV2 the abbreviation corresponding to the statistic test in the equation (2.18) and the one corresponding to $V_2^\star$.

Below, some properties are introduced based on example and simulation studies in [26].

Properties 2.6.0.1.

- *These tests automatically adjust the weighting functions without pre-specifying weights; this implies that these tests are not restricted to only being powerful for specific cases of non-proportional design such as early orlate effect, or cross curves.*
- *For such simulations, V1 and V2 are more powerful than some other statistical tests which might be expected to be more powerful than the logrank test in some non-proportional configurations.*
- *In the simulation studies, V2 performed well when in the presence of early effect, and V1 appeared to be the most powerful in general terms.*

Chapter 3

Control of the type I-error

Previously, we described the logrank test which is the best test when the proportional hazard assumption is respected. We have explained other methods for use in the case of violation of the assumption. In the following, we will compare these methods with the logrank according to the type I-error. The type I-error occurs when the null hypothesis is true but is rejected. The probability of rejecting the null hypothesis given that it is true is called the significance level or the type I-error rate. Let α denote the significance level. Here, we take $\alpha = 0.05$ which means there is 5% risk of incorrectly rejecting the null hypothesis.

In this chapter, we will first describe the simulation design, then discuss the results and finally establish a conclusion.

3.1 Simulation data

Here, three scenarios with two survival curves representing the survival distributions of the treatment and control group are discussed. In the first scenario, the survival functions are exactly the same in both groups. In the second scenario, the two survival curves cross and the true value of the logrank test statistic equals zero. In the third scenario, the two survival curves cross and the expected mean survival time is the same in both groups. Below, the figure shows two survival distributions for the three scenarios.

In the first scenario, both curves follow an exponential distribution with mean 1.8. The first and second scenarios use the same principle to generate data described as follows: the survival distributions are simulated using the principle created by Bender, Augustin and Blettner [27]. Given the hazard function, the simulated survival time is given by

$$T = \Lambda^{-1}[-log(U)], \quad U \sim Unif[0,1], \tag{3.1}$$

where Λ is the cumulative hazard function (see definition 1.1.3).

Then for the treatment distribution in both scenarios I and II, we take the following hazard function

$$\lambda_{Trt}(t) = 0.65 + 2t. \tag{3.2}$$

.

In the second scenario, the hazard function of the control group is given by

$$\lambda_{Ctr}(t) = 0.01 + 4t - 0.7878264t^2 \tag{3.3}$$

where the quadratic coefficient is calculated such that the value of the logrank test statistic equals 0. In the R-code section, the function "find_d" returns the logrank statistic which depends on the value of quadratic coefficient. Then, the function "uniroot" is used to find the value of the quadratic coefficient in order to have the condition that the logrank statistic equals 0.

In the third scenario, the hazard function of the control group is given by

$$\lambda_{Ctr}(t) = 0.1 + 3.567035t \tag{3.4}$$

where the linear coefficient is calculated such that the expected mean survival time is the same in both groups. In the R-code section, the function "find_d" returns the difference of the expected mean survival time in both groups which depends on the value of the linear coefficient. Then, the function "uniroot" is used to find the value of the linear coefficient in order to have the condition that the difference of the expected mean survival time in both group equals 0 i.e. the expected mean survival time is the same in both groups.

The first scenario is created to illustrate the null hypothesis. In order to be as realistic as possible, the second and the third scenarios mimic the patterns from [28].

3.2 Simulation design

For the simulation, 1000 iterations are taken, then the estimated type I-error is calculated as the proportion of 1000 generated random samples in which we reject the null hypothesis using a two-sided test at the 0.05 significance level.

We consider three censoring configurations with 380 events within each of the configurations. This number of events is calculated based on the number of events needed to have a power of 90% to observe a hazard ratio of 0.75.

In the first configuration, we consider no-censoring and the sample size $n = 380$. In the second configuration, we consider small censoring, i.e. 20% of censoring and the sample size $n = 475$. In the third configuration, we consider substantial censoring i.e. 50% of censoring and the sample size $n = 760$. In addition, the duration for admission in the study is one year for the no-censoring configuration while in the censoring configuration it is three years.

3.3 Results

In this section, the survival methods logrank test (LR), restricted mean survival time (RMST), generalised pairwise comparison with $\tau = 0$ (BT0) and $\tau = 1$ (BT1), weighted

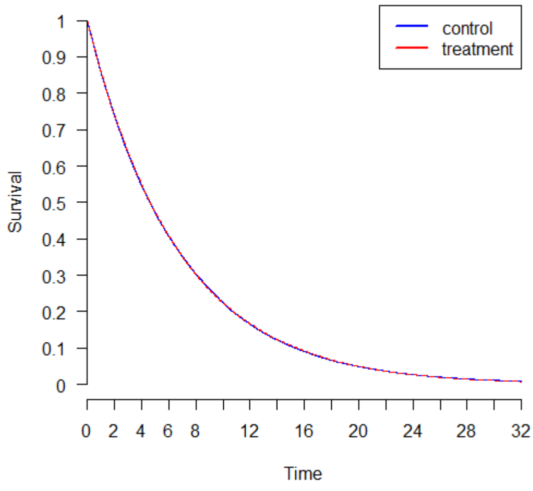


Figure 3.1: First scenario

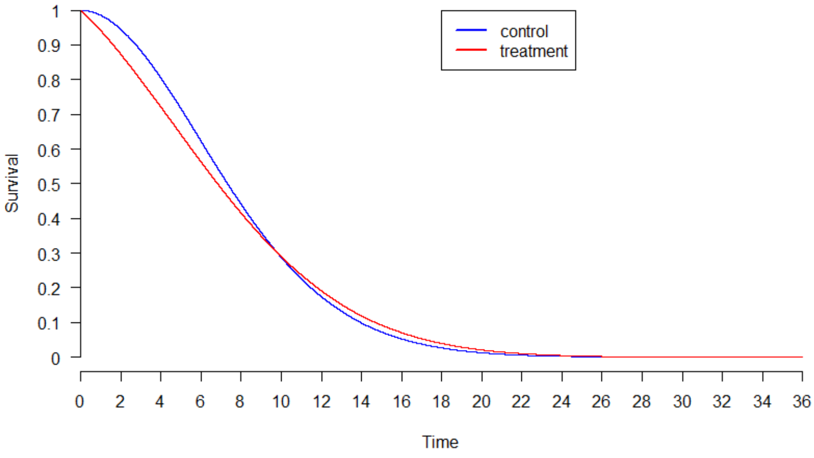


Figure 3.2: Second scenario

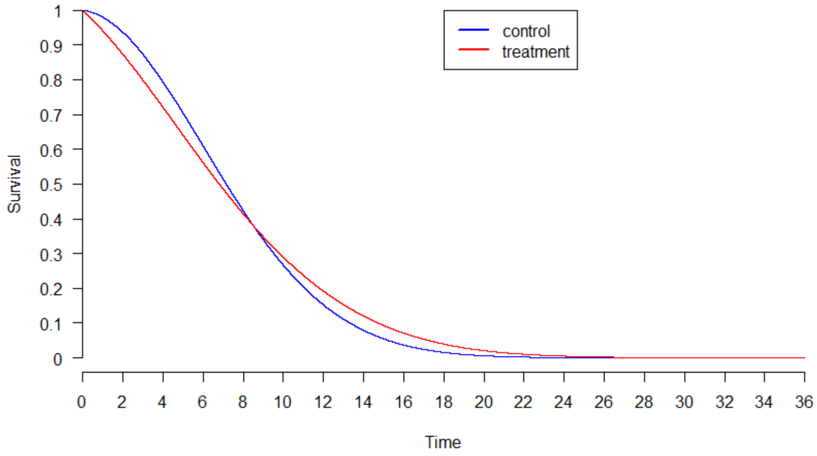


Figure 3.3: Third scenario

logrank test with $(\rho, \gamma) = (0, 1)$ (G01) and $(\rho, \gamma) = (1, 0)$ (G10), adaptive weighted logrank test(AWLR), weighted Kaplan-Meier test using $V_1(c)$ (V1), $V_2(c)$ (V2), the bona fide p-value with $V_1^{\star}(c)$ (BFV1), $V_2^{\star}(c)$ (BFV2) are compared according to the type I-error based on the different scenarios. Table 3.1 shows the simulation results of the estimated type I-error with corresponding 95% confidence interval.

Table 3.1: Type I-error and corresponding 95% confidence interval for each method according to scenario and censoring

Scenario	Cens	LR	RMST	G01	G10	AWLR	BT0	BT1	V1	V2	BFV1	BFV2
I	0%	5.2	5	3.6	6.2	5.4	6.4	4.6	5.7	7	4.9	6.4
		(3.8-6.6)	(3.6-6.4)	(2.4-4.8)	(4.7-7.7)	(4-6.8)	(4.9-7.9)	(3.3-5.9)	(4.3-7.1)	(5.4-8.6)	(3.6-6.2)	(4.9-7.9)
	20%	5.5	5.8	4.7	5.2	5.9	5.7	5.2	6.2	6.5	5.6	5.8
		(4.1-6.9)	(4.4-7.2)	(3.4-6)	(3.8-6.6)	(4.4-7.4)	(4.3-7.1)	(3.8-6.6)	(4.7-7.7)	(5-8)	(4.2-7)	(4.4-7.2)
	50%	5.4	4.9	5.2	5	5.8	4.9	6.3	5.3	5.7	4.9	5.5
		(4-6.8)	(3.6-6.2)	(3.8-6.6)	(3.6-6.4)	(4.4-7.2)	(3.6-6.2)	(4.8-7.8)	(3.9-6.7)	(4.3-7.1)	(3.6-6.2)	(4.1-6.9)
II	0%	4.6	11.1	19.3	20.2	20.4	19.9	4.4	49.5	50.6	45.7	47.7
		(3.3-5.9)	(9.2-13)	(16.9-21.7)	(17.7-22.7)	(17.9-22.9)	(17.4-22.4)	(3.1-5.7)	(46.4-52.6)	(47.5-53.7)	(42.6-48.8)	(44.6-50.8)
	20%	7	11.8	16.8	27.4	30	22.1	4.9	56.3	67.4	51.4	64.4
		(5.4-8.6)	(9.8-13.8)	(14.5-19.1)	(24.6-30.2)	(27.2-32.8)	(19.5-24.7)	(3.6-6.2)	(53.2-59.4)	(64.5-70.3)	(48.3-54.5)	(61.4-67.4)
	50%	7.5	7.8	14.8	21.3	33.6	8.1	6.7	57.1	75	53	72.4
		(5.9-9.1)	(6.1-9.5)	(12.6-17)	(18.8-23.8)	(30.7-36.5)	(6.4-9.8)	(5.2-8.2)	(54-60.2)	(72.3-77.7)	(49.9-56.1)	(69.6-75.2)
III	0%	9.7	5.4	51	9.7	31.9	9.7	14.5	59.9	45.7	55.9	42.7
		(7.9-11.5)	(3.4-6.8)	(47.9-54.1)	(7.9-11.5)	(29-34.8)	(7.9-11.5)	(12.3-16.7)	(56.9-62.9)	(42.6-48.8)	(52.8-59)	(39.6-45.8)
	20%	6.7	6.1	41.4	13.2	30.6	10	17.7	58.5	55	55	53.2
		(5.1-8.2)	(4.6-7.6)	(38.3-44.5)	(11.1-15.3)	(27.7-33.5)	(8.1-11.9)	(15.3-20)	(55.4-61.6)	(51.9-58.1)	(51.9-58.1)	(50.1-56.3)
	50%	4.8	5.3	37.3	8	26.9	6.1	23.7	52.6	58.4	48.4	56.5
		(3.5-6.1)	(3.9-6.7)	(34.3-40.3)	(6.3-9.7)	(24.2-29.6)	(4.6-7.6)	(21.1-26.3)	(49.5-55.7)	(55.3-61.5)	(45.3-51.5)	(53.4-59.6)

3.4 Discussion

In scenario I, where the survival functions are exactly the same in both groups, no rates of type I-error exceed 7%. When there is no censoring, G01, RMST, BT1 and BFV1 better control type I-errors than LR. AWLR and V1 have similar rates to LR and all others have slightly inflated type I-errors. We observe that when there is 20% of censoring, G01 remains the best control rate and all others stay equivalent to the LR except for V1 and V2. When there is 50% of censoring, RMST, BT0 and BFV1 control the type I-error. All others are similar rates to LR except for BT1 which has a slightly inflated type I-error (6.3%). We can conclude that in this scenario, all methods are equivalent to the LR according to the type I-error. We reach the same conclusion as [14]. Indeed, in [14], the authors compared LR with RMST and concluded that both tests have similar type I-errors, whereas in [3], the authors compared LR, G10, G01 and AWLR and concluded that both had similar type I-errors except for AWLR which had inflated type I-error. In our study, AWLR has a similar type I-error to the others. This may be caused by the fact that we do not know on which simulation the authors tested the type I-error. In [2], the authors compare LR, G01, G10 and conclude that G01 has a type I-error which is slightly inflated. However, in our case, G01 has a type I-error close to the nominal level and exceeds it by 0.2% only when there is 50% of censoring. In [26], the authors compare LR, V1 and V2. They show a similar rate to what we have. Indeed, they observed similar type I-error rates for V1 but the rate corresponding to V2 is smaller than the rate that we observed in our study simulation. Furthermore, we can notice that each confident interval for each method in each censored configuration contains the nominal level of 5%. We conclude for this scenario that the LR, RMST, AWLR and BFV1 are relatively conservative, whereas, G01, G10, BT0, V1, BFV2 and V2 statistics gradually approach the nominal level of 5% as the censoring rates increase and BT1 has a slightly inflated type I-error as the censoring rates increase.

In scenario II, at first sight, we notice that the type I-error corresponding to the weighted Kaplan-Meier test (V1, V2, BFV1, BFV2) explodes the type I-error. Indeed, the type I-error rate exceeds 45% which means that in at least 45% of cases, this method rejects the equality of the survival function. In the case of no censoring, the logrank test (LR) correctly controls the type I-error with a rate of 4.6%. However, BT1 has the smallest the type I-error rate, at 4.4%. These two tests are followed by RMST with a slightly inflated type I-error of 11.1%. All other tests have inflated type I-errors that exceed 19%. The more censoring increases, the more type I-errors inflate for LR, AWLR and BT1. LR, RMST and BT1 remain methods which have the lowest type I-errors. In the case of 20% of censoring, BT1 has the smallest type I-error (4.9%) followed by the LR test with 7%. This order remains the same in the case of 50% of censoring. Moreover, in the last case of censoring, LR, RMST, BT0 and BT1 have a similar type I-error rate around 7%. It was expected that LR controls type I-error better than other methods when no censoring occurs because the survival curves were built such that the true value of the logrank test statistic is equal to zero. Moreover, in this scenario, BT0 has the smallest type I-error rate in each censoring configuration. Furthermore, RMST, G10 and BT0 have a similar behaviour. Indeed, when there is 20% of censoring the type I-error grows compared to the case of no censoring whereas when there is 50% of censoring, the type I-error decreases and the rate is smaller than the rate in the case of no censoring for RMST and BT0.

In scenario III, at first sight, we notice that the type I-error corresponding to the weighted Kaplan-Meier test (V1, V2, BFV1, BFV2) and G01 explodes the type I-error. Indeed, the type I-error rate exceeds 42%. As expected, when there is no censoring, RMST controls the type I-error (5.4%) better than other methods because the nominal level 5% is included in the confidence interval. Indeed, in this scenario, the survival curves were built such that the expected mean survival time is the same in both groups. However, LR, G10 and BT0 directly follow the RMST test with 9.7% and AWLR has a type I-error exceeding 30%. When there is 20% of censoring, RMST still controls the type I-error with 6.1% and the nominal level 5% is included in its confidence interval. LR directly follows RMST with 6.7% and the nominal level 5% is almost included in its confidence interval. G10, BT0 and BT1 have a slightly inflated type I-error that exceeds 10%. However, in the case of 50% of censoring, LR, RMST and BT0 better control type I-error (4.8%, 5% and 6.1% respectively) followed by G10 with 8%. All other tests exceed 23%. Furthermore, we notice as in the previous scenario that RMST, G10 and BT0 have a similar behaviour. Indeed, they have an increased type I-error for 20% of censoring as the type I-error decrease at 50% of censoring and the rate is smaller than the rate when no censoring occurs.

To conclude this discussion, the weighted Kaplan-Meier test is not significant to the control of type I-error. As expected in scenarios II and III, LR and RMST correctly control the type I-error in the case of no censoring which is caused by the building of the curves. Moreover, in the second scenario, BT1 remains close to the type I-error rate of LR in each censoring configuration. The more censoring increases, the closer the LR and RMST methods are according to the type I-error.

Chapter 4

Application on real data

In this chapter, we apply the above survival methods to a real data example from [29].

4.1 Resected stage III melanoma

In 2011, Pegylated interferon alfa-2b (PEG-IFN-α-2b) was approved by the US food and Drug administration for the adjuvant treatment[1] of patients suffering from melanoma. The patients have microscopic or gross nodal involvement within 84 days of definitive surgical resection, which also includes complete lymphadenectomy, based on the European Organisation for Research and Treatment of Cancer (EORTC) 18991 trial outcome data. In this trial, EORTC 18991, the treatment group corresponded to patients receiving the adjuvant treatment with PEG-IFN-α-2b. At the beginning there were 1,256 patients with stage III melanoma.

Admissible patients were aged 18 to 70 years with historically documented stage III melanoma after complete regional lymphadenectomy. All patients were aware of whether or not they were receiving the adjuvant treatmen. Patients were randomly selected in a 1:1 ratio to treatment group according to minimization techniques operated by the EORTC data center. Patients were assigned to each group for a duration of 5 years of observation. Table 4.1 shows the number of patients in each group as well as the number of events in each group.

Table 4.1: Baseline characteristics

	PEG-IFN-alpha-2b group	Control group
Randomized	627	629
Number of events	384	406

[1] Adjuvant therapy is an additional cancer treatment given after primary treatment in order to reduce the risk of the cancer coming back.

Figure 4.1 illustrates the relapse-free survival curves of patients from PEG-IFN-α-2b and one from a control group. In [29], the assumption of proportional hazards was checked and it was indicated that the curves were not proportional. Furthermore, Figure 4.1 shows that the late effect is present which has already detailed in chapter 2. Indeed, the curves are closed at the beginning and then are separated. This is what we usually observe in immunotherapy.

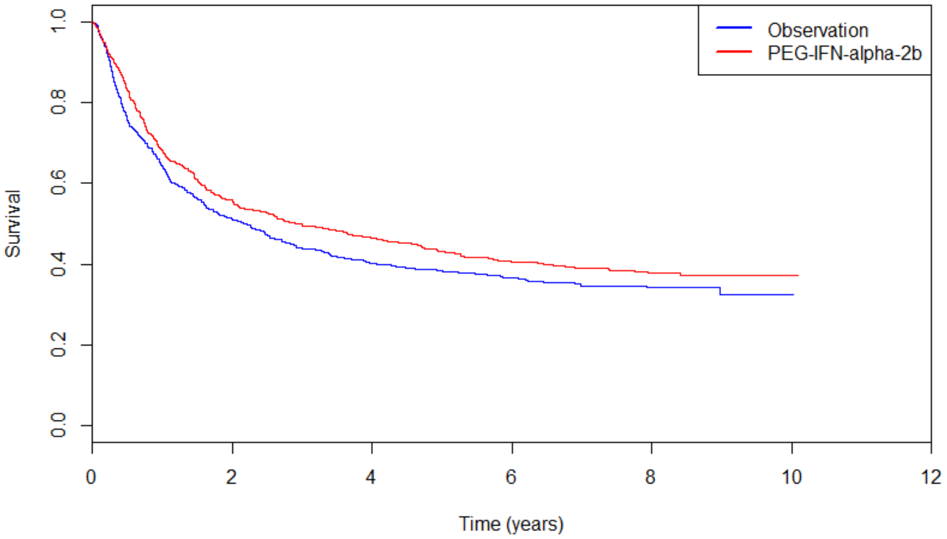


Figure 4.1: Survival comparison of overall population of relapse-free survival

4.2 Results

Now, we apply all methods to the real dataset. The corresponding p-value of the different methods developed in chapter 2 are listed in Table 4.2.

Firstly, it appears that the choice of approach has implications for whether or not a statistically significant effect is concluded. We also notice that G01 tends to reject the null hypothesis (p-value=0.344) whereas G10 identifies a significant difference between the two groups (p-value=0.026). This can be explained by the fact that G10 gives more weight to departure which occurs late in time whereas G01 gives more weight to early departure. In our case, we have a departure which occurs late in time, so G10 is more appropriate. As expected according to the literature [13] and [30], the p-value of the LR and RMST tests are closed and these methods do not reject the null hypothesis. Moreover, BFV1 does not provide evidence of a difference in treatment effect. However, at a 5% significance level, AWLR, BT0, BT1, V1, V2 and BFV2 conclude that there are significant difference between the two groups. The literature [16] confirms that in their example, LR failed to show a significant survival benefit while BT showed a survival benefit.

Table 4.2: EORTC 18991:P-values from the different methods

LR	RMST	G01	G10	AWLR	BT0	BT1	V1	V2	BFV1	BFV2
0.055	0.051	0.344	0.026	0.046	0.035	0.05	0.049	0.011	0.053	0.012

Chapter 5

Conclusion

In cancer research, the proportional hazard assumption is often violated; particularly in immunotherapy, because a pattern of delayed treatment effect is observed. As this assumption is in doubt, the most popular statistical test, the logrank test, loses power, and so other statistical methods have been developed in the literature which do not depend on this assumption.

In this master's thesis, we first fixed the background of survival analysis with the definition of survival function and hazard function. We demonstrated the main estimator of survival function, known as the Kaplan-Meier estimator. Before describing the statistical methods used in the simulation, we established the aim of testing the comparison of survival curves of two groups, treatment versus control and the different types of patterns which can appear when non-proportional hazard is observed. After that, the logrank test was described and illustrated using the Gehan data.

Furthermore, five other statistical methods were described: restricted mean survival time, generalised pairwise comparison, weighted logrank, adaptive weighted logrank and weighted Kaplan-Meier tests. The restricted mean survival time difference is defined as the difference in the area under the estimation of the survival function of treatment and control groups. As the name of the method implies, this difference of area is restricted with a pre-specified threshold. This threshold is either specified before the study or defined as the minimum of the largest observed time in each of the two groups. The main advantage of this method is the interpretability of this method from a clinical perspective such as loss of life expectancy. Generalised pairwise comparison is an uncommon method to test the null hypothesis. Indeed, this method takes pairs of individuals from each group, treatment and control, and classifies them into four groups: "Favourable", "Unfavourable", "Neutral" and "Uninformative". From this we use a pairwise score which can be seen as the probability of a random patient in the treatment group of having a "better outcome" than a random patient in the control group minus the opposite probability. Then we obtain the net benefit, i.e. the proportion in favour of treatment, which allows testing of the null hypothesis. Weighted and adaptive weighted logranks are an updating of the logrank method. The first one uses the Fleming Harrington weight. This weight can be adapted according to the patterns of the survival curves. Indeed, according to the parameter of the weight we can add more weight to early departure or more weight to departure which occurs late in time. On the other hand, adaptive weight is based on the model of Yang and Prentice. This new model provides a more accurate description of the data in certain non proportional hazard situations. Finally, the weighted Kaplan-Meier test is the weighting of the area of the difference between the Kaplan-Meier estimator of the survival curves from treatment and control groups up to a certain threshold defined by the study. Four different statistics based on the different weighted Kaplan-Meier estimators were described.

Then, we compared these six methods according to the type I-error in the simulation study. We discussed three different simulation scenario. In the first scenario, the survival function was exactly the same in both groups. In the second scenario, the two survival curves crossed and the true value of the logrank test statistics equalled zero. In the third scenario, the two survival curves crossed and the expected mean survival time was the same in both groups. In each scenario, three censoring configurations $(0\%, 20\%, 50\%)$ were investigated. For each pattern, 1000 random samples were generated. From these simulation studies, we noted that in the first scenario, regardless of the censoring rate, all type I-errors were close to 5%. For scenario II, LR and BT1 correctly controlled the type I-error which was expected for LR due to the structure of scenario II. The same observation was made in scenario III for LR, RMST, G10 and BT0, and RMST was expected to be the best method to control the type I-error because this scenario was built on this method. For other situations, the type I-error rate was highly inflated for V1, V2, BFV1 and BFV2.

Finally, an application to real data was conducted. The methods were applied to compare survival data from the EORTC phase III clinical trial comparing adjuvant therapy with pegylated-interfon-α-2b with an observation groups in stage III melanoma patients. The relapse-free survival figures and results of the comparison between the two treatment groups were given. The conclusion drawn from the results on this dataset was that the choice of the approach has implications on whether or not a statistically significant effect is reached.

At this stage of this master's thesis, we cannot affirm that there is one method which is adapted for the case of a non proportional hazard situation. A perspective to continue this master's thesis is to compare all methods, except that related to the weighted Kaplan-Meier test which had a highly inflated type I-error rate, according to the type II-error. This test of power had already been done to compare some methods in [3] and [2] but only for crossed curves. The simulation for the type II-error could be curves which are more realistic in an immunotherapy context, i.e. curves which illustrate a delayed effect. Moreover, after investigating the power of the methods we could focus on the quality of life in patients. Another perspective is to include methods such as a combined test [8] or a cure model [7] which were not included in this master's thesis due to the lack of R-function.

Appendix A

Other methods to estimate the restricted mean survival time

Method 1: Pseudo-observations

We will first define pseudo-observations in a general manner and then apply them to the case of restricted mean survival time.

Pseudo-observations (or pseudo-values) are non-parametric leave-one-out methods which enable estimation of the RMST [31].

Let $X_1, ..., X_n$ be i.i.d. copies of a random variable X and $\theta = \theta(X)$ be a parameter of the form

$$\theta = E[\phi(X)] = \int \phi(x) \, \mathrm{d}F_X(x),$$

where $\phi(.)$ is some function of X.

Assume that we have, an unbiased estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$ of θ based on the entire sample $\mathbf{X}= \{X_1, ..., X_n\}$. Then,

$$E[\hat{\theta}] = \int \hat{\theta}(\mathbf{x}) \, \mathrm{d}F_X(x) = \theta.$$

Definition A.0.1. Let $X_1, ..., X_n$ be the i.i.d. random variables and let $\hat{\theta}(\mathbf{X})$ be an unbiased estimator of the parameter θ defined as before. For each X_j the *pseudo-observation* is defined by

$$\hat{\theta}_j(\mathbf{X}) = n\hat{\theta}(\mathbf{X}) - (n-1)\hat{\theta}^{-j}(\mathbf{X}), j = 1, ..., n, \tag{A.1}$$

where $\hat{\theta}^{-j}(.)$ is the estimate based on the sample without the observation j.

We can use the previous definition to estimate the restricted mean survival time μ_τ. Let us begin by defining the function $\phi(.)$ and the parameter θ. The function $\phi(.)$ is given by $\phi(X) = min(X, \tau)$, and the parameter θ by μ_τ.
Using definition 2.2.1, μ_τ may be estimated by

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)\mathrm{d}t,$$

where $\hat{S}(t)$ is the Kaplan-Meier estimator (see definition 1.2.1). Then the j^{th} pseudo-observation is given by

$$\begin{aligned}\hat{\mu}_{\tau j} &= \int_0^{\tau} \hat{S}_j(t)\,\mathrm{d}t \\ &= n\int_0^{\tau} \hat{S}(t)\,\mathrm{d}t - (n-1)\int_0^{\tau} \hat{S}^{-j}(t)\,\mathrm{d}t \quad j = 1, ..., n,\end{aligned}$$

where $\hat{S}^{-j}(t)$ is the pseudo-observation[1] given by

$$\hat{S}_j(t) = n\hat{S}(t) - (n-1)\hat{S}^{-j}(t),$$

where $\hat{S}^{-j}(.)$ is the Kaplan-Meier estimator of $S(.)$ based on the observations $i \neq j$. Then,

$$\hat{\mu}_{\tau} = \frac{1}{n}\sum_{j=1}^{n} \hat{\mu}_{\tau j},$$

$\hat{\mu}_{\tau}$ is an unbiased estimator for the RMST when the Kaplan-Meier estimate is an unbiased estimator of the survival function[2].

This method is implemented in R using package "pseudo" and the function "pseudomean" to model the survival function based on the restricted mean.

Method 2: Flexible parametric survival model

To begin, the spline function [31] should be defined in order to estimate the cumulative hazard function.

Definition A.0.2. Let [a,b] be an interval on $\mathbb{R}$ and let ν=$\{\nu_1, ..., \nu_r\}$ be a real number, called *knots*, satisfying

$$a = \nu_0 < \nu_1 < ... < \nu_r < b = \nu_{r+1}$$

A function $s_{\nu} : [a, b] \to \mathbb{R}$ is a *spline function of order d* if the following are satisfied:

- $s_{\nu}(t)$ is a polynomial of order d on each interval $[\nu_i, \nu_{i+1}], i = 0, ..., r$.
- $s_{\nu} \in C^{d-1}[a, b]$ i.e. the spline $s_{\nu}(.)$ and its derivatives up to $d-1^{th}$ order are continuous at all points in the interval $[a, b]$ and in particular at all knots ν.

[1] we take as $\phi(X_j) = \mathbb{I}(X_j > t)$, $j = 1, ..., n$ and the parameter $\theta = S(.)$ evaluated at time t

[2] This happens when the censored survival time is, independent of participants' covariates, within treatment groups.

Royston and Parmar [32] proposed approximation of the baseline log cumulative hazard function using the restricted cubic spline function[3].

First, the log of the cumulative baseline hazard $\Lambda_0(t)$ is approximated using a function of the log of time

$$\ln \Lambda_0(t) = \gamma_0 + \gamma_1 s_1(\ln t) + ... + \gamma_{K_0+1} s_{K_0}(\ln t)$$

where $\gamma_i (i = 0, ..., K_0 + 1)$ are the regression parameters and $s_i (i = 1, ..., K_0)$ is the i^{th} spline basis function.

Here, K_0 denotes the number of distinct internal knots, which is the join-point in log time of a pair of adjacent cubic polynomial segments.

The restricted mean survival time can be written as

$$\hat{\mu}_\mathcal{T} = \int_0^\mathcal{T} S(t)\,\mathrm{d}t = \int_0^\mathcal{T} exp(-\Lambda(t))\,\mathrm{d}t.$$

Taking $\ln \Lambda_0(t) = s(\ln t|\boldsymbol{\gamma}, K_0)$, then the log cumulative hazard function is given by

$$\ln \Lambda(t) = s(\ln t|\boldsymbol{\gamma}, K_0) + s(\ln t|\boldsymbol{\delta}, K_1)x + \beta x.$$

where x presents the treatment arm indicator. The additional term $s(\ln t|\boldsymbol{\delta}, K_1)x$ is taken in account for the non-proportional hazards. The parameter $\boldsymbol{\gamma} = (\gamma_0, ..., \gamma_{K_0+1})$ is the regression coefficient in the baseline spline function which has K_0 knots; the $\boldsymbol{\delta} = (\delta_1, ..., \delta_{K_1})$ is the regression coefficient in the interaction spline function which has K_1 knots. As the model is parametric, the parameters can be estimated by the maximum likelihood approach.

The number of knots increasing implies that the model complexity is also increased. Usually, 3 degrees of freedom are used for the baseline distribution i.e. $K_0 = 2$.

This method is implemented in R using package "flexsurv" and the function "rmst_ generic" is a generic function to find the restricted mean of a distribution.

[3] Restricted cubic splines are splines that are restricted to be linear before the first knot and after the last knot.

Appendix B

Delta Method

This appendix is based on [33]. The delta method is a method which allows the user to derive the variance of a function of an asymptotically normal random variable with known variance.

The delta method is defined as follows:

Definition B.0.1. Let $X_1, ..., X_n$ be random variables with mean μ and variance σ^2. If $\sqrt{n}[X_n - \mu] \to N(0, \sigma^2)$ then for all functions g such that $g'(\mu)$ exist and is non-zero value, we have

$$\sqrt{n}[g(X_n) - g(\mu)] \to N(0, \sigma^2(g'(\mu))^2).$$

Appendix C

Derivation of the formula in Table 2.4

For each element in Table 2.4, we use the general formula:

$$s_{ij} = \mathbb{P}(X_i^0 > Y_j^0 + \tau | X_i, Y_j, X_i^0 \geq X_i, Y_j^0 \geq Y_j) - \mathbb{P}(Y_j^0 > X_i^0 + \tau | X_i, Y_j, X_i^0 \geq X_i, Y_j^0 \geq Y_j) \tag{C.1}$$

No censoring: $(\epsilon_i, \eta_j) = (1, 1)$

In favour of treatment,

$$\begin{aligned} \mathbb{P}(X_i^0 > Y_j^0 + \tau | X_i, Y_j, X_i^0 = X_i, Y_j^0 = Y_j) &= X_i > Y_j + \tau, \\ &= \begin{cases} 1 & \text{if } x_i > y_j + \tau \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

In favour of control,

$$\begin{aligned} \mathbb{P}(Y_j^0 > X_i^0 + \tau | X_i, Y_j, X_i^0 = X_i, Y_j^0 = Y_j) &= Y_j > X_i + \tau, \\ &= \begin{cases} 1 & \text{if } y_j > x_i + \tau \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Then, we get the first row,

$$s_{ij} = \begin{cases} 1 & \text{if } x_i - y_j > \tau \\ -1 & \text{if } x_i - y_j < -\tau \\ 0 & \text{otherwise.} \end{cases}$$

Only X_i is censored : $(\epsilon_i, \eta_j) = (0, 1)$

In favour of treatment,

$$\mathbb{P}(X_i^0 > Y_j^0 + \tau | X_i, Y_j, X_i^0 > X_i, Y_j^0 = Y_j) = \mathbb{P}(X_i^0 > Y_j + \tau | X_i, Y_j, X_i^0 > X_i),$$

by applying the conditional probabilities formula,

$$= \frac{\mathbb{P}((X_i^0 > Y_j + \tau) \cap (X_i^0 > X_i)|X_i, Y_j)}{\mathbb{P}(X_i^0 > X_i|X_i, Y_j)},$$

using definition 1.1.1,

$$\begin{aligned}
&= \frac{\mathbb{P}(X_i^0 > max(Y_j + \tau, X_i)|X_i, Y_j)}{S_{Trt}(X_i)},\\
&= \frac{S_{Trt}(max(Y_j + \tau, X_i))}{S_{Trt}(X_i)}\\
&= \begin{cases} \dfrac{S_{Trt}(X_i)}{S_{Trt}(X_i)} & \text{if } x_i > y_j + \tau \\ \dfrac{S_{Trt}(Y_j + \tau)}{S_{Trt}(X_i)} & \text{otherwise,} \end{cases}\\
&= \begin{cases} 1 & \text{if } x_i > y_j \tau \\ \dfrac{S_{Trt}(Y_j + \tau)}{S_{Trt}(X_i)} & \text{otherwise.} \end{cases}
\end{aligned}$$

In favour of control,

$$\begin{aligned}
\mathbb{P}(Y_j^0 > X_i^0 + \tau|X_i, Y_j, X_i^0 > X_i, Y_j^0 = Y_j) &= \mathbb{P}(Y_j > X_i^0 + \tau|X_i, Y_j, X_i^0 > X_i)\\
&= 1 - \mathbb{P}(X_i^0 > Y_j - \tau|X_i, Y_j, X_i^0 > X_i)
\end{aligned}$$

applying the previous result,

$$\begin{aligned}
&= 1 - \begin{cases} 1 & \text{if } x_i > y_j - \tau \\ \dfrac{S_{Trt}(Y_j - \tau)}{S_{Trt}(X_i)} & \text{otherwise,} \end{cases}\\
&= \begin{cases} 0 & \text{if } x_i > y_j - \tau \\ 1 - \dfrac{S_{Trt}(Y_j - \tau)}{S_{Trt}(X_i)} & \text{otherwise.} \end{cases}
\end{aligned}$$

Then we get the second row,

$$s_{ij} = \begin{cases} 1 & \text{if } x_i - y_j > \tau \\ \dfrac{S_{Trt}(Y_j + \tau) + S_{Trt}(Y_j - \tau)}{S_{Trt}(X_i)} - 1 & \text{if } x_i - y_j < -\tau \\ \dfrac{S_{Trt}(Y_j + \tau)}{S_{Trt}(X_i)} & \text{otherwise.} \end{cases}$$

Remark C.0.0.1. Because the survival function is a non-negative decreasing function,

$$S_{Trt}(Y_j + \tau) < S_{Trt}(X_i) \text{ when } |x_i - y_j| < \tau, \text{ then } 0 \leqslant s_{ij} \leqslant 1$$

and

$$S_{Trt}(Y_j - \tau) < S_{Trt}(X_i) \text{ and } S_{Trt}(Y_j + \tau) < S_{Trt}(X_i) \text{ when } x_i - y_j < -tau, \text{ then } -1 \leqslant s_{ij} \leqslant 1.$$

Only Y_j is censored : $(\epsilon_i, \eta_j) = (1, 0)$

As treatment and control groups work symmetrically, we get the same formula for s_{ij} as in the previous case by exchanging X_i^0, X_i, S_{Trt} and Y_j^0, Y_j, S_{Ctr}.

X_i and Y_j are censored : $(\epsilon_i, \eta_j) = (0, 0)$

In favour of treatment, when $X_i > Y_j + \tau$

by applying the law of total probabilities,

$$\begin{aligned}
&\mathbb{P}(X_i^0 > Y_j^0 + \tau | X_i, Y_j, X_i^0 > X_i, Y_j^0 > Y_j) \\
&= \mathbb{P}((X_i^0 > Y_j + \tau) \cap (X_i > Y_j^0 + \tau) | X_i, Y_j, X_i^0 > X_i, Y_j^0 > Y_j) \\
&\quad + \mathbb{P}((X_i^0 > Y_j + \tau) \cap (X_i < Y_j^0 + \tau) | X_i, Y_j, X_i^0 > X_i, Y_j^0 > Y_j),
\end{aligned}$$

by applying conditional probabilities formula,

$$\begin{aligned}
&= \mathbb{P}(X_i > Y_j^0 + \tau | X_i, Y_j, X_i^0 > X_i, Y_j^0 > Y_j) \\
&\quad + \frac{\mathbb{P}(X_i^0 > T_j^0 + \tau) \cap (X_i < Y_j^0 + \tau) \cap (X_i^0 > X_i) \cap (Y_j^0 > Y_j) | X_i, Y_j)}{\mathbb{P} X_i^0 > X_i, Y_j^0 > Y_j | X_i, Y_j}, \\
&= 1 - \frac{S_{Ctr}(X_i - \tau)}{S_{Ctr}(Y_j)} + \frac{\mathbb{P}((X_i^0 > Y_j^0 + \tau) \cap (X_i < Y_j^0 + \tau) | X_i, Y_j}{S_{Trt}(X_i) S_{Ctr}(Y_j)},
\end{aligned}$$

by using $\mathbb{P}(A > B) = -\int_{-\infty}^{\infty} S_A(s) \mathrm{d}S_B(s)$ where $A \perp\!\!\!\perp B$,

$$\begin{aligned}
&= 1 - \frac{S_{Ctr}(X_i - \tau)}{S_{Ctr}(Y_j)} - \frac{\int_{X_i - \tau}^{\infty} \mathbb{P}(X_i^0 > t + \tau) \mathrm{d}\mathbb{P}(Y_j^0 > t)}{S_{Trt}(X_i) S_{Ctr}(Y_j)}, \\
&= 1 - \frac{S_{Ctr}(X_i - \tau)}{S_{Ctr}(Y_j)} - \frac{\int_{X_i - \tau}^{\infty} S_{Trt}(t + \tau) \mathrm{d}S_{Ctrt}(t)}{S_{Trt}(X_i) S_{Ctr}(Y_j)}.
\end{aligned}$$

When $X_i < Y_j + \tau$ i.e. $Y_j > X_i - \tau$, by applying the law of total probabilities,

$$\begin{aligned}
&\mathbb{P}(X_i^0 > Y_j^0 + \tau | X_i, Y_j, X_i^0 > X_i, Y_j^0 > Y_j) \\
&= \frac{\mathbb{P}((X_i^0 > Y_j^0 + \tau) \cap (X_i^0 > X_i) \cap (Y_j^0 > Y_j) | X_i, Y_j)}{\mathbb{P}(X_i^0 > X_i, Y_j^0 > Y_j | X_i, Y_j)} \\
&= \frac{\mathbb{P}((X_i^0 > Y_j^0 + \tau) \cap (X_i^0 > X_i) \cap (Y_j^0 > Y_j) | X_i, Y_j)}{\mathbb{P}(X_i^0 > X_i | X_i)\mathbb{P}(Y_j^0 > Y_j | Y_j)} \\
&= \frac{\mathbb{P}((X_i^0 > Y_j^0 + \tau) \cap (X_i^0 > X_i) \cap (Y_j^0 > Y_j) | X_i, Y_j)}{S_{Trt}(X_i) S_{Ctr}(Y_j)}.
\end{aligned}$$

As $(X_i^0 > Y_j^0 + \tau) \cap (Y_j^0 > Y_j) \Rightarrow X_i^0 > Y_j + \tau$ and $(X_i^0 > Y_j^0 + \tau) \cap (Y_j > X_i - \tau) \Rightarrow X_i^0 > X_i$ then $(X_i^0 > Y_j^0 + \tau) \cap (X_i^0 > X_i) \cap (Y_j^0 > Y_j) = (X_i^0 > Y_j^0) \cap (Y_i^0 > Y_j)$,

$$= \frac{\mathbb{P}((X_i^0 > Y_j^0 + \tau) \cap (Y_j^0 > Y_j) | X_i, Y_j)}{S_{Trt}(X_i) S_{Ctr}(Y_j)},$$

by using $\mathbb{P}(A > B) = -\int_{-\infty}^{\infty} S_A(s) \mathrm{d} S_B(s)$ where $A \perp\!\!\!\perp B$,

$$\begin{aligned}
&= -\frac{\int_{Y_j}^{\infty} \mathbb{P}(X_i^0 > t + \tau) \mathrm{d}\mathbb{P}(Y_j^0 > t)}{S_{Trt}(X_i) S_{Ctr}(Y_j)}, \\
&= -\frac{\int_{Y_j}^{\infty} S_{Trt}(t + \tau) \mathrm{d} S_{Ctr}(t)}{S_{Trt}(X_i) S_{Ctr}(Y_j)}.
\end{aligned}$$

In favour of control, as treatment and control groups work symmetrically, we get the same formula as in the previous case by exchanging X_i^0, X_i, S_{Trt} and Y_j^0, Y_j, S_{Ctr}.

When $Y_j > X_i + \tau$ i.e. $X_i < Y_j - \tau$,

$$\begin{aligned}
&\mathbb{P}(Y_j^0 > X_i^0 + \tau | X_i, Y_j, X_i^0 > X_i, Y_j^0 > Y_j) \\
&= 1 - \frac{S_{Trt}(Y_j - \tau)}{S_{Trt}(X_i)} - \frac{\int_{Y_j - \tau}^{\infty} S_{Ctr}(t + \tau) \mathrm{d} S_{Trt}(t)}{S_{Ctr}(Y_j) S_{Trt}(X_i)}.
\end{aligned}$$

When $Y_j < X_i + \tau$ i.e. $X_i > Y_j - \tau$,

$$\begin{aligned}
&\mathbb{P}(Y_j^0 > X_i^0 + \tau | X_i, Y_j, X_i^0 > X_i, Y_j^0 > Y_j) \\
&= \frac{-\int_{X_i}^{\infty} S_{Ctr}(t + \tau) \mathrm{d} S_{Trt}(t)}{S_{Ctr}(Y_j) S_{Trt}(X_i)}.
\end{aligned}$$

Then we get the last row,

$$s_{ij} = \begin{cases} \mathbb{P}(X_i^0 > Y_j^0 + \tau|\textbf{.}, X_i > Y_j + \tau) - \mathbb{P}(Y_j^0 > X_i^0 + \tau|\textbf{.}, Y_j < X_i + \tau) \\ \mathbb{P}(X_i^0 > Y_j^0 + \tau|\textbf{.}, X_i < Y_j + \tau) - \mathbb{P}(Y_j^0 > X_i^0 + \tau|\textbf{.}, Y_j > X_i + \tau) \\ \mathbb{P}(X_i^0 > Y_j^0 + \tau|\textbf{.}, X_i < Y_j + \tau) - \mathbb{P}(Y_j^0 > X_i^0 + \tau|\textbf{.}, X_i \in [Y_j - \tau, Y_j < X_i + \tau]) \end{cases}$$

$$= \begin{cases} 1 - \dfrac{S_{Ctr}(X_i - \tau)}{S_{Ctr}(Y_j)} + \dfrac{-\int_{X_i-\tau}^{\infty} S_{Trt}(t+\tau)\mathrm{d}S_{Ctr}(t) + \int_{X_i}^{\infty} S_{Ctr}(t+\tau)\mathrm{d}S_{Trt}(t)}{S_{Trt}(X_i)S_{Ctr}(Yj)} & \text{if } x_i - y_j > \tau \\[2ex] -1 + \dfrac{S_{Trt}(Y_j - \tau)}{S_{Trt}(X_i)} + \dfrac{-\int_{Y_j}^{\infty} S_{Trt}(t+\tau)\mathrm{d}S_{Ctr}(t) + \int_{Y_j-\tau}^{\infty} S_{Ctr}(t+\tau)\mathrm{d}S_{Trt}(t)}{S_{Trt}(X_i)S_{Ctr}(Yj)} & \text{if } x_i - y_i < -\tau \\[2ex] \dfrac{\int_{X_i}^{\infty} S_{Ctrt}(t+\tau)\mathrm{d}S_{Trt}(t) - \int_{Y_j}^{\infty} S_{Trt}(t+\tau)\mathrm{d}S_{Ctr}(t)}{S_{Trt}(X_i)S_{Ctr}(Y_j)} & \text{otherwise} \\ \end{cases}$$

Remark C.0.0.2. When X_i and Y_j are censored, the following result is used to derive the formula:

$$\mathbb{P}(A > B) = -\int_{-\infty}^{+\infty} S_A(s)\mathrm{dS_B(s)}, \quad A \perp\!\!\!\perp B.$$

Let us prove this equality:

By the Tower porperty we have,

$$\begin{aligned} \mathbb{P}(A > B) &= E[\mathbb{P}(A > B|B)], \\ &= \int_{-\infty}^{+\infty} \mathbb{P}(A > B|B = s)\mathrm{dF_B(s)}, \end{aligned}$$

by the independency of the variable,

$$\begin{aligned} &= \int_{-\infty}^{+\infty} \mathbb{P}(A > s)\mathrm{dF_B(s)}, \\ &= \int_{-\infty}^{+\infty} S_A(s)\mathrm{dF_B(s)}, \\ &= -\int_{-\infty}^{+\infty} S_A(s)\mathrm{dS_B(s)}. \end{aligned}$$

Appendix D

Derivation of the limite 2.12

We express the equation 2.9 as follows

$$\lambda_{Trt} = \frac{\theta_1\theta_2\lambda_{Ctr}}{\theta_1(1-S_{Ctr})+\theta_2 S_{Ctr}}. \tag{D.1}$$

Thus,

$$1-\frac{\lambda_{Ctr}}{\lambda_{Trt}} = \frac{\theta_1(1-S_{Ctr})+\theta_2 S_{Ctr}}{\theta_1\theta_2}.$$

Moreover, we have

$$\theta_1 = \exp(\gamma_1\theta) \text{ and } \theta_2 = \exp(\gamma_2\theta).$$

Then the limit becomes,

$$\begin{aligned}&\lim_{\theta\to 0}\frac{(1-\frac{\lambda_{Ctr}}{\lambda_{Trt}})}{\theta}\\ &=\lim_{\theta\to 0}\frac{\theta_1(1-S_{Ctr})+\theta_2 S_{Ctr}}{\theta_1\theta_2\theta}\end{aligned}$$

by L'Hôpital's rule, we get

$$\begin{aligned}&=\lim_{\theta\to 0}\frac{(\gamma_1+\gamma_2)\exp((\gamma_1+\gamma_2)\theta)-\gamma_1\exp(\gamma_1\theta)(1-S_{Ctr})-\gamma_2\exp(\gamma_2\theta)S_{Ctr}}{(\gamma_1+\gamma_2)\exp((\gamma_1+\gamma_2)\theta)\theta+\exp((\gamma_1+\gamma_2)\theta)}\\ &=\gamma_1+\gamma_2-\gamma_1(1-S_{Ctr})-\gamma_2 S_{Ctr}\\ &=\gamma_2(1-S_{Ctr})+\gamma_1 S_{Ctr}\\ &=\Omega.\end{aligned}$$

Appendix E

Supplementary material of weighted Kaplan-Meier test

E.1 Choice of c fix at 1.65

Let us fix c at 1.65, under the null hypothesis, since $Z(t) \sim N(0,1)$ then $W_c(t) \simeq 1.65$ for most of $t \in [0, \varsigma]$. This implies that the distribution of $V_1(c)$ which can be considered as a linear combination of dependent standard normal random variables, would not have a long right tail. Moreover, under an alternative hypothesis, for a large observed $Z(t)$, i.e. $Z(t) > 1.65$, $W_c(t) = Z(t)$, the observed $V_1(c)$ would also be large.

E.2 General approach to choose c adaptively

Let assume that we can generate a good approximation to the distribution of the process $V_1(c)$ indexed by $c \in [0, \eta]$ under the null hypothesis.

Let $v_1(c)$ be the realisation of $V_1(c)$ and its p-value $p(c)$ which is calculated using the approximation of the distribution of $V_1(c)$ under the null hypothesis. We define the most significant $p(c)$ in $c \in [0, \eta]$ by $p_b = \min\{p(c) : c \in [0, \eta]\}$. Note that a small value of p_b would strengthen the alternative hypothesis. We want to choose the threshold value c in order to have a statistical significance based on p_b. This means that we need to find the null distribution of the random part of p_b, $P_b = \min\{P(c) : c \in [0, \eta]\}$, where $P(c) = S_{V_1(c)}(V_1(c))$ and $S_{V_1(c)}(v)$ is the survival function of $V_1(c)$. Using the standard martingale theory and the central limit theorem, it can be shown that $Z(\cdot)$ converges in distribution to a limiting Gaussian process $G(\cdot)$ [34]. Uno et al. show in [26] that $V_1(c)$ and $P(c)$, as a process in c, converge weakly to $\phi(c) = \int_0^\xi \max\{G(t), c\}G(t)\mathrm{d}t$ and $U(c) = S_{\phi(c)}(\phi(c))$ respectively.

A perturbation-re-sampling method can be used to empirically approximate the limiting distribution under the null hypothesis. Therefore, the distribution of the process $(\hat{S}_i(t) -$

$S_i(t)), i = \{T, C\}$ can be approximated by following

$$Q_i(t) = -\hat{S}_i(t) \sum_{j=1}^{n_i} \left[\frac{1}{\sum_{k=1}^{n_i} \mathbb{I}(x_{ik} \geqslant x_{ij})} \delta_{ij} \, \mathbb{I}(x_{ij} \leqslant t) \xi_{ij} \right], \tag{E.1}$$

where x_{ij} is the observed value of X_{ij}, δ_{ij} is the censored indicator and $\{\xi_{ij}, i \in \{Trt, Ctr\}, j = 1, ..., n_i\}$ is a random sample from a distribution with a mean equal to 0 and a variance equal to 1.

Bibliography

[1] TM. Therneau P. Grambsch. Proportional hazards test and diagnostics based on weighted residuals. *Biometrika*, 10(2):408–422, 1982.

[2] Yamen Hou Huilin Chen Zheng Chen Huimin Li, Dong Han. Statistical inference method for two crossing survival curves: a comparison of method. *Plos one*, 10(1):1–18, january 2015.

[3] Bart Spiessens Andrea Callegaro. Testing treatment effect in randomized clinical trials with possible non-proportional hazards. *Statistical in biopharmaceutical research*, 9:204–211, Janvier 2017.

[4] Institut de Cancérologie du CHU. Quelques mots sur l'immunothérapie, 2010.

[5] D. Oakes D.R. Cox. *Analysis of survival data.* London, 1988.

[6] Delphine Maucort-Boulch Janez Stare. Odds ratio, hazard ratio and relative risk. *Metodoloski zvezki*, 13(1):59–67, 2016.

[7] Megan Othus et al. Cure models as a useful statistical tool for analysing survival. *Clinical Can*, 18(14):3731–3735, 2012.

[8] M. K. Parmar Patrick Royston. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Medical R*, 16(16), 2016.

[9] Touati Nathan Michal Kicinski. Statistical challenges in immunotherapy trials. 2017.

[10] S. Chevret C. Alberti, J.-F.Timsit. Analyse de survie : le test du logrank. *Mémento biostatistique*, 2005.

[11] Mahesh K.B. Parmar Patrick Royston. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*, 30:2409–2421, May 2011.

[12] Jayne F. Tierney Mahesh K.B. Parmar Yinghui Wei, Patrick Royston. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. *Statistics in Medicine*, 34:2881–2898, June 2015.

[13] Mahesh KB Parmar Patrick Royston. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, 13(152), 2013.

[14] Qiang Xu Xun Lin. A new method for the comparison of survival distributions. *Pharmaceutical Statistics*, 9:67–76, 2009.

[15] Uno H Solomon SD Pfeffer MA Schindler JS Wei LJ Zhao L, Tian L. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials*, 9(4):570–577, 2012.

[16] Marc Buyse. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, 29:3245–3257, 2010.

[17] Brie Ozenne Laurent Roche Pascal Roy Julien Peron, Marc Buyse. An extension of generalized pairwise comparisons for prioritized outcomes in the precense of censoring. *Statistical Methods in Medical Research*, 0(0):1–10, 2016.

[18] Thomas R. Fleming Monika Peckova. Adaptive test for testing the difference in survival distributions. *Lifetime Data Analysis*, 9:223–238, 2003.

[19] Seung-Hwan Lee. On the versatility of the combination of the weighted log-rank statistics. *Computational Stastistics and data analysis*, 51:6557–6564, March 2007.

[20] Davi. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566, 1982.

[21] David P. Harrington Thomas R. Fleming. Counting processes and suvival analysis. *Wiley*, 1991.

[22] Ryszard Zieliťnski Agnieszka Rossa. A simple improvement of the kaplan-meier estimator.

[23] Ross Prentice Song Yang. Improved logrank-type test for survival data using adaptive weights. *Biometrics*, 66((1)):30–38, March 2010.

[24] Song Yang and Ross Prentice. Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika Trust*, 92(1):1–17, 2005.

[25] TR Fleming MS Pepe. Weighted kaplan-meier statistics: a class of distance test for censored survival data. *Biometrics*, 45:497–507, 1989.

[26] Brian Claggett Hajime Uno, Lu Tian and L.J. Wei. A versatile test for equality of two survival functions based on weighted differences of kaplan-meier curves. *stat. med.*, 34(28):3680–3695, December 2015.

[27] Maria Blettner Ralf Bender, Thomas Augustin. Generating survival time to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, February 2005.

[28] Eugene D kwon et al. Ipilimunab versus placebo after radiotherapy in patients with metastatic castration-resistant prostate cancer that had progressed after docetaxel chemotherpy (ca184-043): a multicentre, randomised, double-blind, phase 3 trial. *Lancet oncol*, 15:700–712, 2014.

[29] Alexander M.M. Eggermont et al. Long-term results of the randomized phase iii trial EORTC 18991 of adjuvant therapy with pegylated interferon alfa-2b versus observation in restricted stage III melanoma. *Journal of clinical oncology*, 30(31):3810–3818, november 2012.

[30] Hajime Uno et al. Moving beyond the hazard ratio in quatifying the between-group difference in suvival analysis. *Journal of cli*, 32(22):2380–2384, August 2014.

[31] Rikke Nørmark Mortensen. Pseudo-observations in survival analysis. Master thesis, Aalborg University, 2013.

[32] Parmar MK. Royston P. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197, 2002.

[33] Annie Wang Alex Gold, Nat Olin. What is the delta method., 2018.

[34] Richard Gill. Large sample behaviour of the product-limit estimator on the whole line. *The annals of statistics*, 11(1):49–58, 1983.

[35] Tom Brody. *Clinical trials: Study design, endpoints and biomarkers, drug safety A, FDA and ICH guidelines*. Elsevier, 2012.

[36] Julien Peron. Generalized pairwise comparison on immuno-oncology clinical trial data : a case study. 2017.

R-code

```
install.packages("survival")

library(survival)


install.packages("survRM2")

library(survRM2)


install.packages("YPmodel")

library(YPmodel)


install.packages("PwrGSD") # use version 3.4 for R.

library(PwrGSD)


install.packages("survAWKMT2")

library(survAWKMT2)


http://cran.r-project.org/bin/windows/Rtools/ # install Rtools34.exe

install.packages("devtools")


devtools::install_github("bozenne/BuyseTest") # version 1.3.2

library(BuyseTest)


install.packages("doSNOW")

library(doSNOW)
```

SIMDATARAND

Input:
- n: number of observations.
- acrdur: duration of the admission in the study.

Output: The function returns a list which contains the treatment assignment, and time of accrual. 0 for control group and 1 for treatment group. The accrual of time is the moment where the patient starts the study.

```
SIMDATARAND <- function(n,acrdur){

      acrtime <- runif(n,0,acrdur)

      trt <- rbinom(n,1,0.5)

 return(list (trt=trt,acrtime=acrtime))

}
```

hazard_trt

Input: t : time.

Output: The function returns the value of the hazard function at the specific time t.

```
hazard_trt <- function(t){

   0.65+2*t

}
```

hazard_ctr

Input:
- t: time.
- d: quadratic coefficient.

Output: The function returns the value of the hazard function at the specific time t.

- For lrs0

```
hazard_ctr <- function(t,d){

  0.01+4*t+d*t^2

}
```

- For mst0

```
hazard_ctr <- function(t,d){

  0.1+d*t
}
```

SIMSURV.FROM.HAZARDS

Input:
- lambda: hazard function.
- n: number of observations.
- maxtime: At least the longest time in analysis.
- mintime: Minimum time in analysis. Default 0.

Output: The function returns the survival time for each observation.

```
SIMSURV.FROM.HAZARDS <- function (lambda,n,maxtime,mintime=0){
  S <- function (time)
  {
    S <- exp (-integrate (lambda,lower=0,upper=time)[[1]])
    S
  }
  fun_to_minimize <- function (time,d)
  {
    S <- exp (-integrate (lambda,lower=0,upper=time)[[1]])
    S-u
  }
  time <- rep(NA,n)
  for (i in 1:n)
  {
    u <- runif(1)
    if  (u<S (maxtime))
   {
      time[i]<-maxtime
    }
    else
    {
      time[i]<-uniroot (fun_to_minimize,lower=mintime,upper=maxtime)[[1]]
    }
  }
  return (time=time)

}
```

SIMSURV.FROM.HAZARDS.CTR

Input:
- lambda: hazard function.
- n: number of observations.
- maxtime: At least the longest time in analysis.
- mintime: Minimum time in analysis. Default 0.
- d: argument to be passed to lambda.

Output: The function returns the survival time for each observation.

```
SIMSURV.FROM.HAZARDS.CTR <- function (lambda,n,maxtime,mintime=0,d){

  S <- function (time,d)

  {

    S <- exp (-integrate (lambda,lower=0,upper=time,d=d)[[1]])

    S

  }

  fun_to_minimize <- function (time,d)

  {

    S <- exp (-integrate (lambda,lower=0,upper=time,d=d)[[1]])

    S-u

  }

  time<-rep(NA,n)

  for (i in 1:n)

  {

    u <- runif(1)

    if (u<S(maxtime,d=d))

    {

      time[i] <- maxtime

    }

    else

    {

      time[i] <-
uniroot(fun_to_minimize,lower=mintime,upper=maxtime,d=d)[[1]]

    }
  }

  return (time=time)

}
```

find_d

Input:
- d: the argument to be found
- trt: the group indicator. The element of this vector takes either 1 for treatment group or 0 for control group.
- effect: "lrs0" if the argument d is such that the true value of the logrank test statistic equals 0. "mst0" if the argument d is such that the mean survival time is the same in both groups.

Output: The function returns either the logrank test statistic or the difference of mean survival time depending on the argument d.

```
find_d <- function (d,trt,effect){

  n <- length(trt)

  ntrt <- sum(trt)

  nco <- n-sum(trt)

  time <- rep(NA,n)

  event <- rep(1,n)

  timeco <- c()

  timetrt <- c()

  for (j in 1:ntrt){

    timetrt[j] <-
SIMSURV.FROM.HAZARDS(lambda=hazard_trt,n=1,maxtime=5,mintime=0)

  }

  for (j in 1:nco){

    timeco[j]<-
SIMSURV.FROM.HAZARDS.CTR(lambda=hazard_ctr,n=1,maxtime=5,mintime=0,d=d)

  }

  time[which(trt==0)] <- timeco

  time[which(trt==1)] <- timetrt

  if(effect=="lrs0"){

    fit <- survdiff(Surv(time,event)~trt)

    return (fit$exp[1]-fit$obs[1])

  }
  if(effect=="mst0"){
    tau <- min(max(time[which(trt==1)]),max(time[which(trt==0)]))
    fit <- rmst2(time=time, status=event, arm=trt,tau=tau)
    return (fit[[5]][1,1])
  }
}
```

The value of the quadratic coefficient depending on the effect

```
d <- uniroot(find_d,interval=c(-1,0),trt,effect)[[1]]
```

- For "lrs0" : d = -0.7878264

```
d <- uniroot(find_d,interval=c(3,4),trt,"mst0")[[1]]
```

- For "mst0": d = 3.567035

simsurv_curve

Input:
- trt: the group indicator. The element of this vector takes either 1 for treatment group or 0 for control group.
- effect: "lrs0" if the argument d is such that the true value of the logrank test statistic equals 0. "mst0" if the argument d is such that the mean survival time is the same in both groups.

Output: The function returns a data frame containing a vector "trt" for treatment assignment, a vector of "time" for survival time and a vector of "event" containing only 1.

```
simsurv_curve<-function(trt,effect){
  n<-length(trt)
  ntrt<-sum(trt)
  nco<-n-sum(trt)
  time<-rep(NA,n)
  event<-rep(1,n)
  timeco<-c()
  timetrt<-c()
  if(effect=="lrs0"){

      d<-uniroot(find_d,interval=c(-1,0),trt,effect)[[1]]

      print(d)

  }


  if(effect=="mst0"){

      d<-uniroot(find_d,interval=c(3,4),trt,effect)[[1]]

      print(d)

  }


  for(j in 1:ntrt){

    timetrt[j]<-
SIMSURV.FROM.HAZARDS(lambda=hazard_trt,n=1,maxtime=5,mintime=0)

  }


  for (j in 1:nco){

    timeco[j]<-
SIMSURV.FROM.HAZARDS.CTR(lambda=hazard_ctr,n=1,maxtime=5,mintime=0,d=d)

  }

  time[which(trt==0)]<-timeco

  time[which(trt==1)]<-timetrt

  return(data.frame(trt=trt,time=time,event=event))

}
```

simsurv_zero

Input: trt: the group indicator. The element of this vector takes either 1 for treatment group or 0 for control group.

Output: The function returns a data frame containing a vector "trt" for treatment assignment, a vector of "time" for survival time and a vector of "event" containing only 1.

```
simsurv_zero<-function(trt){

   n<-length(trt)

   ntrt<-sum(trt)
   nco<-n-ntrt

   time<-rep(NA,n)
   event<-rep(1,n)
   timeco<-c()
   timetrt<-c()

   for (j in 1:n){
     time[j]<-rexp(n=1,rate=1.8)
   }
   return(data.frame(trt=trt,time=time,event=event))
 }
```

The curve of simulation data

```
datarand<-SIMDATARAND(n=10^6,acrdur=1)
## Figure 3.1 ##
data<-simsurv_zero(datarand$trt)

## Figure 3.2##
data<-simsurv_curve(datarand$trt,effect = "lrs0")

## Figure 3.3##
data<-simsurv_curve(datarand$trt,effect = "mst0")

## plot each curve ##

time_month<-data$time*12
fit<-survfit(Surv(time_month,event)~trt,data=data,conf.type="none")
plot(fit,col=c("blue","red"),xlim=c(0,36),axes=FALSE,ylab="Survival",xlab="
Time")
axis(side=2,at=seq(0,1,0.1),labels=seq(0,1,0.1),las=2)
axis(side=1,at=seq(0,36,2),labels=seq(0,36,2),las=0)
```

```
legend("topright",legend=c("control","treatment"),lty=1,lwd=2,col=c("blue",
"red"))
```

Control of type I-error

Hazard function

```
hazard_lrs0<-function(t)

{

  0.01+4*t-0.7878264*t^2

}


hazard_mst0<-function(t)

{

  0.1+3.567035*t

}
```

Simsurv

Input:

- trt: the group indicator. The element of this vector takes either 1 for treatment group or 0 for control group.
- nevents: number of events.
- acrtime: time of accrual.
- effect: "lrs0" if the argument d is such that the true value of the logrank test statistic equals 0. "mst0" if the argument d is such that the mean survival time is the same in both groups.

Output: The function returns a data frame containing a vector "trt" for treatment assignment: 1=treatment group and 0=control group, a vector of "time" for survival time and a vector of "event" containing the status indicator, 1=dead and 0=censored.

```
simsurv<-function(trt,nevents,acrtime,effect){

  n<-length(trt)

  ntrt<-sum(trt)

  nco<-n-sum(trt)

  event<-rep(1,n)

  timeco<-c()
  timetrt<-c()

  if(effect=="zero"){
    time<-c()
    for (j in 1:n){
      time[j]<-rexp(n=1,rate=1.8)
    }
  }

  if(effect=="lrs0"){
    time<-rep(NA,n)
    for(j in 1:ntrt){
      timetrt[j]<-
SIMSURV.FROM.HAZARDS(lambda=hazard_trt,n=1,maxtime=5,mintime=0)
    }
    for (j in 1:nco){
      timeco[j]<-
SIMSURV.FROM.HAZARDS(lambda=hazard_lrs0,n=1,maxtime=5,mintime=0)
    }
    time[which(trt==0)]<-timeco
    time[which(trt==1)]<-timetrt
  }

  if(effect=="mst0"){
    time<-rep(NA,n)
    for(j in 1:ntrt){
      timetrt[j]<-
SIMSURV.FROM.HAZARDS(lambda=hazard_trt,n=1,maxtime=5,mintime=0)
    }
    for (j in 1:nco){
      timeco[j]<-
SIMSURV.FROM.HAZARDS(lambda=hazard_mst0,n=1,maxtime=5,mintime=0)
    }
    time[which(trt==0)]<-timeco
    time[which(trt==1)]<-timetrt
}

  realtime<-time+acrtime

  censoredobs<-cumsum(event[order(realtime)])>nevents

  event[order(realtime)][censoredobs]<-0

  return(data.frame(trt=trt,time=time,event=event))

}
```

FITMODELS

Input:
- data: data frame in which to interpret the variable occurring in the model function.
- b: number of iterations.

Output: The function fits models for one sample and returns the 2-sided p-value for each model.

```
FITMODELS<-function(data,b){

  ### Log-rank test ###

  print("logrank")

  fit<-survdiff(Surv(time, event) ~ trt, data = data)

  z<-(fit$obs[1]-fit$exp[1])/fit$var[1,1]^0.5

  lrt2s<-1-pchisq(fit$chisq, length(fit$n)-1)


  ### Restricted mean survival time ###

  print("rmst")

  tau<-min(max(data$time[which(data$event==1 &
data$trt==1)]),max(data$time[which(data$event==1 & data$trt==0)]))

  fit<-rmst2(time=data$time, status=data$event, arm=data$trt,tau=tau)

  z<-fit[[5]][1,1]/(fit$RMST.arm0$rmst.var+fit$RMST.arm1$rmst.var)^0.5

  drmst2s<-rmst2(time=data$time, status=data$event,
arm=data$trt,tau=tau)[[5]][1,4]


  ### Weigthed logrank test ###

  ## G01 ##

  print("g01")

  fit2s<-wtdlogrank(Surv(time, event) ~ trt, data=data,WtFun = "FH",  param
= c(0, 1),sided=2)
```

```
  U<-fit2s$stat
  V<-fit2s$var
  wlr01<-1-pchisq(U^2/V,1)


  ## G10 #
  print("g10")
  fit2s<-wtdlogrank(Surv(time, event) ~ trt, data=data,WtFun = "FH",param =
c(1, 0),sided=2)
  U<-fit2s$stat
  V<-fit2s$var
  wlr10<-1-pchisq(U^2/V,1)


  ### Adaptively weighted log-rank test ###
  print("awlrt")
  data1<-as.data.frame(cbind(V1=data$time,V2=data$event,V3=data$trt))
  adlrt <- YPmodel.adlgrk(data=data1)$pval # two-sided #A function to
calculate p-value of the adaptive weighted logrank test.


  ### Generalised pairwise comparison tau=0 ###
  data3<-as.data.frame(cbind(time=data$time,event=data$event,trt=data$trt))
  print("bt0")
  BT_tau0<-
BuyseTest(data=data3,treatment="trt",endpoint="time",type="timeToEvent",thr
eshold=as.numeric(0),censoring="event",method.tte="Peron",n.resampling=1000
,cpus=8)
  BT2s_0<-summary(BT_tau0)$table$p.value[1]


  ### Generalised pairwise comparison tau=1 ###
  print("bt1")
  BT_tau1<-
BuyseTest(data=data3,treatment="trt",endpoint="time",type="timeToEvent",thr
eshold=as.numeric(1),censoring="event",method.tte="Peron",n.resampling=1000
,cpus=8)
  BT2s_1<-summary(BT_tau1)$table$p.value[1]


  ### A versatile test based on weighted differences of KM curves ###
  print("wkm")
  data2<-
as.data.frame(cbind(time=data$time,status=data$event,arm=data$trt))

  tau = max(data2[data2[,2]==1,1])

  val = AWKMT2(indata=data2, tau=tau, nmethod=1000, seed=b*123+123,
v1=TRUE, v2=TRUE, test="2_side")

  wkmt<-c(val[[2]],val[[3]],val[[4]],val[[5]])



return(list(twosided=c(lrt2s,drmst2s,wlr01,wlr10,adlrt,BT2s_0,BT2s_1,wkmt))
)

}
```

RUNSIM

Input:
- settings: settings containing number of observations, number of events, effect, duration of admission in the study.
- B: number of iterations.
- fn: character string.

Output: The function executes functions of the different models and returns a table containing the p-value of each model for each iteration.

```
RUNSIM<-function(settings,B,fn){

  for (s in 1:length(settings[,1])){

    pvalues2s<-matrix(nrow=0,ncol=12)

    colnames(pvalues2s)<-
c("b","lrt","drmst","wlr01","wlr10","adlrt","BT_0","BT_1","crude_V1","crude
_V2","bona_fide_V1","bona_fide_V2")



    for (b in 1:B){

      print(s)

      print(b)

      datarand<-SIMDATARAND(n=settings$n[s],acrdur=settings$acrdur[s])

      data<-
simsurv(trt=datarand$trt,nevents=settings$nevents[s],acrtime=datarand$acrti
me[s],effect=settings$effect[s])

      results<-FITMODELS(data,b)

      rm(datarand)

      rm(data)

      pvalues2sb<-c(b,results$twosided)
      pvalues2s<-rbind(pvalues2s,pvalues2sb)
    }
    results2s<-cbind(settings[s,][c(rep(1,B)),],pvalues2s)


    if (s>1){
      results2sold<-read.csv2(file=paste(fn,"2s.csv",sep=""))
      results2s<-rbind(results2sold,results2s)
    }
    write.csv2(results2s,file=paste(fn,"2s.csv",sep=""),row.names=FALSE)
  }
}
```

TABLE

Input: fn: character string.

Output: The function returns a table for the control of type I-error of each model.

```
TABLE<-function(fn){
  results2s<-read.csv2(file=paste(fn,"2s.csv",sep=""))
  settings<-unique(results2s[,1:4])
  S<-length(settings[,1])
  size2s<-matrix(nrow=S,ncol=11)
  for (s in 1:S){
    results2s_s<-results2s[which(results2s[,1]==settings[s,1] &
results2s[,2]==settings[s,2] & results2s[,3]==settings[s,3] &
results2s[,4]==settings[s,4]),]


    B<-length(results2s_s[,1])
    size2s[s,]<-apply(results2s_s[,c(6:16)]<0.05,2,sum)/B
  }
  table2s<-cbind(settings,size2s)
  colnames(table2s)<-c(colnames(settings),colnames(results2s[6:16]))
  write.csv2(table2s,file=paste(fn,"table2s.csv",sep=""),row.names=FALSE)
}
```

Table 3.1

```
settings<-
data.frame(n=rep(c(380,475,760),3),nevents=rep(380,9),effect=c(rep("zero",3
),rep("lrs0",3),rep("mst0",3)),acrdur=rep(c(1,3,3),3))

RUNSIM(settings = settings,B=1000,fn="simulation1")

TABLE(fn="simulation1")
```