ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ

Adaptive Clinical Trial designs for survival outcomes testing the proportionality of hazards assumption

Χαράλαμπος Δημητρίου Σταυρόπουλος

ΕΡΓΑΣΙΑ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος

Συμπληρωματικής Ειδίκευσης στη Στατιστική

Πλήρους Παρακολούθησης (Full-time)

Αθήνα

Σεπτέμβριος 2018

ΕΥΧΑΡΙΣΤΙΕΣ

I would like to thank my supervisor Dr. Dimitris Karlis for his guidance and support that provided me throughout my work on this thesis. Also, I would like to thank Dr. Urania Dafni for her valuable advice which helped me completing my work. Finally, I would like to thank Frontier Science Foundation Hellas for providing me with a scholarship for this thesis.

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

I am Charalampos Stavropoulos, born in 1993 and I have graduate from the Department of Mathematics of the National and Kapodistrian University of Athens. I am currently a graduate student at Athens University of Economics and Business. I attend the MSc in Statistics. Frontier Science Foundation Hellas has provided me with a scholarship for conducting my master thesis. Also, the main results of my master thesis were presented at XXIXth International Biometric Conference, Barcelona, Spain, 8-13 July 2018, under the title: “Adaptive Clinical Trial designs for survival outcomes testing the proportionality of hazards assumption” with co-authors Dr. Dimitris Karlis and Dr. Urania Dafni.

ABSTRACT

Charalampos Stavropoulos

Adaptive Clinical Trial designs for survival outcomes testing the proportionality of hazards assumption

September 2018

At this thesis we will examine a widely used assumption in the context of clinical trials, the assumption of proportional hazards. This assumption is about the relationship between the hazard functions of the groups that participate in the trial. More specifically, it is assumed that the ratio of hazard functions is constant through time. Based on that assumption, the researchers can use the Log Rank test and estimate the sample size that is needed. However, this is a very strict assumption and in this thesis we investigate its impact on the trial when it does not hold, in terms of power and sample size estimation. As an alternative, we investigate the Restricted Mean Survival Time (RMST), which is the mean survival time up to a certain point. We will use the difference between RMSTs for testing the difference between survival groups and we will compare this method to the Log Rank test under cases of proportional and non-proportional hazards. The comparison will be in terms of sample size estimation and power. Finally, we will provide a new clinical trial design that will start with the Log Rank test and at a certain point will test the proportionality assumption. If the assumption is rejected, the trial will adapt to testing the difference between RMSTs. The adaptive design's performance will be compared to that of a simple design which does not test the proportionality assumption and uses the Log Rank test. The comparison will be in terms of sample size estimation and power.

ΠΕΡΙΛΗΨΗ

Χαράλαμπος Σταυρόπουλος

Προσαρμοστικοί σχεδιασμοί κλινικών δοκιμών για αποτελέσματα επιβίωσης που ελέγχουν την υπόθεση αναλογικότητας των κινδύνων

Σεπτέμβριος 2018

Σε αυτή τη διπλωματική εργασία θα εξετάσουμε μια υπόθεση που χρησιμοποιείται πολύ συχνά στο πλαίσιο των κλινικών δοκιμών. Η υπόθεση αφορά τη σχέση μεταξύ των συναρτήσεων κινδύνου των ομάδων που συμμετέχουν στη κλινική δοκιμή. Η υπόθεση είναι ότι ο λόγος των συναρτήσεων κινδύνου είναι σταθερός στο χρόνο. Με την υπόθεση αυτή οι ερευνητές μπορούν να χρησιμοποιήσουν τον έλεγχο Log Rank και να εκτιμήσουν το απαιτούμενο μέγεθος δείγματος. Αυτή η υπόθεση όμως είναι πολύ αυστηρή και σε αυτή την εργασία θα ερευνήσουμε την επίδραση που έχει στην εκτίμηση του μεγέθους δείγματος και την ισχύ, όταν χρησιμοποιείται λανθασμένα. Σαν εναλλακτική, θα ερευνήσουμε τον περιορισμένο μέσο χρόνο επιβίωσης (RMST), που είναι ο μέσος χρόνος επιβίωσης μέχρι ένα σημείο. Θα χρησιμοποιήσουμε τη διαφορά μεταξύ των RMST, για να ελέγξουμε τη διαφορά στις ομάδες επιβίωσης και θα συγκρίνουμε αυτή τη μέθοδο με τον έλεγχο Log Rank σε περιπτώσεις αναλογικότητας και μη-αναλογικότητας των συναρτήσεων κινδύνου. Η σύγκριση θα αφορά την εκτίμηση μεγέθους δείγματος και την ισχύ. Στο τέλος θα παρουσιάσουμε ένα σχεδιασμό κλινικής δοκιμής που θα ξεκινάει με τον έλεγχο Log Rank και σε ένα συγκεκριμένο σημείο θα ελέγχει την υπόθεση αναλογικότητας. Αν η υπόθεση απορριφθεί, ο σχεδιασμός θα προσαρμοστεί στο να ελεγχθεί η διαφορά των RMST. Η απόδοση του προσαρμοστικού σχεδίου θα συγκριθεί με αυτή ενός απλού σχεδιασμού που χρησιμοποιεί τον έλεγχο Log Rank και δεν ελέγχει την αναλογικότητα. Η σύγκριση θα γίνει στην εκτίμηση μεγέθους δείγματος και την ισχύ.

ΚΑΤΑΛΟΓΟΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

CHAPTER 1: Introduction

Clinical trials have greatly contributed to the process of developing new treatments and clinical research in general, since their establishment. New interventions that are developed can be tested and prove their superiority against existing ones for the same diseases. The possible adverse effects and other relative matters (e.g. what should be the dose of a new drug in order to be efficient but not toxic?) are also examined. Before the clinical trials become a part of clinical research, the new interventions were tested in questionable way and their efficiency was estimated without strict and well established statistical methods. Nowadays, there are fundamental principles for conducting a clinical trial and very strict regulations. The reason is that the outcome of the trial will either approve or not approve the use of a new treatment. So it is very important that the new treatments that get the approval have well established their efficiency. Furthermore, since the clinical trials are testing new treatments to people, they must be conducted in a way that meets very strict ethical standards. The target of the trial must be clear from the beginning and the trial must not continue without a sufficient scientific reason. A general rule is that the possible adverse effects of the new intervention must not exceed the possible benefits that the participants can get.

Since the clinical trials are comparing different treatments based on data they need statistical methods. Biostatisticians have developed many methods for estimating the probability of death at specific time intervals, the life expectancy and other aspects that interest the clinicians. There also many statistical methods for comparing different treatments and show their superiority (or non-inferiority). However, statistical testing provides results with the probability of two errors (type I and type II error) and the probability of each error must be pre-specified at the beginning of the trial. The number of the participants (sample size), the duration of the trial and the statistical method for comparing the treatments and many other factors must also be pre-specified. Biostatisticians make statistical assumptions about the nature of the treatment’s effect on participants in order to specify those factors. When designing a clinical trial, researchers should pre-specify all those factors in an efficient way that protects the participants, examines sufficiently the new treatment’s effect and also keep the duration, the sample size and the cost in reasonable levels. For that purpose the researchers can use flexible and adaptive designs.

The purpose of this thesis is to investigate a widely used assumption in the context of clinical trials. The assumption is that of the proportionality between the hazard functions of the treatments that are compared. The hazard function is the probability of someone to die in a very small time interval (it is explained in more details at Section 1.4). So the assumption is basically that the ratio of those values between patients that get different treatments is constant. This assumption is very convenient because it enables the use of a widely used model presented from Cox (1972), the Cox model. This model provides easy interpretation for the comparison between treatments and a very useful statistical context for analyzing them. Also it enables the use of the Log Rank test which is widely used in clinical trials and it tests the hazard ratio. However, the assumption is very strict, and in practice it holds rarely. This thesis investigates the impact that the use of the assumption has on the designing of the trial when it does not hold.

Another measure that can be used in the survival analysis is the Restricted Mean Survival Time (RMST). It is the mean survival time of a patient until a certain time point. If we get the difference of the RMSTs from two different groups that get different treatments, computed at the same time point, we can construct a test that compares the two groups. This test does not need an assumption about the relationship of the hazard functions and because of that, in this thesis we will compare its performance against the Log Rank test in various cases of non-proportional hazards. Also the two tests will be compared under the proportionality case as well, in order to have a complete investigation on what problems can the RMST test solve and when we can replace the Log Rank test with that.

The final purpose of the thesis is to use both tests and construct an adaptive design that takes the advantage of both of them. Since the use of the Log Rank test is sometimes not the best choice, the adaptation will replace it with the RMST test. The adaptive design must be very strict on under what conditions, the replacement happens and what else must change in order not to stray from the initial target of the clinical trial. Also the adaptation will be based on the results from the comparison between the two tests, in order to get the more suitable one for each case. After the construction of the adaptive design, we will investigate its performance and compare it to the performance of a typical design that just uses the Log Rank test for all cases. Specifically, we are interested in comparing the two designs in terms of power and sample size needed and to investigate if we can get benefits from the use of the adaptation. The comparisons will be conducted through simulations of different cases of non-proportional hazards. Now we will give a brief presentation of the context of each chapter.

In the first chapter (Survival Analysis), basic concepts of the survival analysis will be presented such as: the survival distribution, the hazard function and the Cox Model. These concepts are crucial for understanding the clinical trials and the proportionality assumption.

After the survival analysis basics, the thesis presents the basics of the clinical trials. It presents the definition and various types of clinical trials and the concepts of power and sample size calculation. In the second chapter (Clinical Trials), the impact of the wrong use of the proportionality assumption on the sample size calculation is investigated through a simulation study.

The third chapter (Restricted Mean Survival Time) presents the definition of the RMST and also the test of the RMST difference. Then, it compares the sample size that the RMST difference test needs with the one that the Log Rank test needs, for identical cases. The comparisons are done with a simulation study.

The fourth chapter (A New Adaptive Design) presents analytically the adaptation that aims to fix the problems caused by the violation of the proportionality assumption. There is also a simulation study that examines different cases of violation and how the adaptive design performs under them. Its performance is compared to the design that only uses the proportionality assumption and to the Log Rank test, in terms of sample size and power.

CHAPTER 2: Survival Analysis

2.1 Introduction

Survival analysis is the analysis of time-to-event data. It is the process where we observe the time until an event that we are interested in happens to the subjects of the research. Event can be the death of a patient, or a cancer metastasis, even the breakdown of a machine. Thus, time-to-event data need two time points that can be calculated in many ways (days, months, years etc.). A starting point, where the observation of the object starts and an ending point, where an event happens to the observed object. Their difference is the survival time of interest.

A very common problem that occurs in survival analysis trials is that we do not always observe the full time to the event of interest. For example, in a clinical trial we might want to observe the death (the event of interest) of a patient that receives an experimental drug. However this patient might move out of town and cannot visit the research center anymore, or even die by an irrelevant to his/her illness cause. This observation is not complete, because we did not observe his death. Also in many cases there is a time limitation to the trial. When the time limitation is reached, there may be objects that have not experienced the event of interest yet. All those observations are called censored observations.

In the context of survival analysis, in most of the times we will have censored observations. But a censored observation is not exactly a missing observation, because we know that at least, the event that we are interested in, had not occurred until we stop observing the subject. So if we just remove the observation, like we would do in other statistical context, we will ignore a part of the information from the data. In clinical trials there is always the problem of the sample size, so we cannot afford to lose information. Especially in cases of small period trials an important amount of observation can be censored. The models that are used in survival analysis for describing the different aspects are taking into account both the censored and the uncensored observation. We will describe those models and its purpose later.

The main purpose of survival analysis is to construct a “survival profile” of the populations that are studied, meaning to quantify different aspects of their life expectancy and then compare it to other populations. Next we will describe the hazard function and the survival distribution that helps us achieve this purpose.

2.2 Censored data

As discussed in the introduction, there are many different cases for a censored observation. Three common types are:

1) Right censored data: When we observe a subject until a time point, but then we lose it and we do not observe the event of interest. It is called right censor, because at the time axis we cannot see the right tail.
2) Left censored data: It is when we enroll a subject to a trial, but it has already experienced the event that we are interested in.
3) Interval censored data: When we observe an event, but we do not know the exact time. This can happen if we observe the subject at different time points and we detect that the event has happened. We are not sure about the exact time, but we know that it happened between the last observation and the one before.

It is clear now, that the data are complex. As a result we need a notation that shows the length of the observation, as well as the status of the subject. By status we mean and indicator, that shows if an event has happened or not. Let X_i be the failure time of the i_{th} subject of the data set and c_i the end of its observation. Then the observation will be:

$$(T_i, \delta_i),$$

here $T_i = \min(X_i, c_i)$ and

$$\delta_i = \begin{cases} 1 \; if \; X_i \leq c_i \\ 0 \; if \; X_i > c_i \, . \end{cases}$$

Obviously, if $\delta_i = 1$ the observation is uncensored and hence we have the full information about the event time, and if $\delta_i = 0$ is censored.

Now we will give an example to show how important is to take into account the censored observations. We will simulate a trial where we have 90 patients and we are interested to see their survival times. Their survival function (we will describe it later in details) is Weibull (0.7, 0.9) and their censoring distribution is Weibull (2, 2.9).The survival function gives the probability at each time point that the patient will not experience the event until that point. The censoring distribution gives the probability at each time point that the patient will not become censored until that time point. Obviously if the censoring status changes before the death of the patient, the observation is considered to be censored. This combination leads to 20% censored observations on average.

The trials target is to estimate the survival function (we will describe it later) and to understand how “possible” is for a patient from this population to die at any time point. We will use two different methods. Firstly, we will only use the uncensored observations for the analysis. Then we will use both the censored and the uncensored observations. For the estimation we will use the Kaplan – Meier estimator, a very widely used estimator that will be presented in details later in this chapter and it estimated the survival function. Figure 1 will show the difference of the two methods.

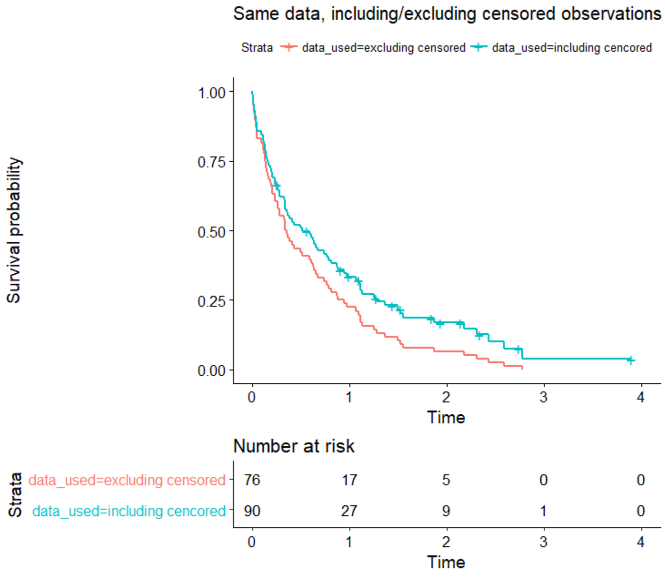


Figure 1: Kaplan-Meier estimators for a dataset without its censored observation and with them.

Figure one shows the Kaplan Meier estimations of the same data set, without removing the censored observations (blue curve) and with having removed them (red curve). At each time point the curve estimates the probability of a patient that he is alive until that time point. So it is clear that the dataset without the censored observations underestimates this probability, since the red curve is lower than the blue one. That happens because as we mentioned before, censored observations inform us that the patient had not experienced the event until he left the trial. So if we ignore that information and just exclude him from the estimation, we do underestimation. Furthermore, note that the blue curve was estimated from 90 people, while the red one was estimated with 76 people. So including censored observations also has an impact on the sample size.

2.3 Survival Function

Let t be the real survival time of a subject. Then t is an observation from a random variable T, with probability density function (p.d.f.) $f(t)$ and cumulative distribution function (cdf) given by:

$$F(t) = P(T \leq t) = \int_o^t f(u)du. \tag{2.1}$$

The Survival function is:

$$S(t) = P(T > t) = 1 - F(t). \tag{2.2}$$

It denotes the probability of observing a survival time greater than a given time t. We need to estimate this function from the data taking into account also the censored observations. A very common method proposed by $Kaplan\,and\,Meier$ (1958) is the Kaplan-Meier estimator. Next we will describe how to calculate this estimator.

Let n be the sample size. And let r be the observed events at the end of the trial. Obviously $r \leq n$. Then let $t(1), \ldots., t(r)$ be the times of the observed events and $t(1) < t(2) < \cdots. < t(r)$ (assuming no ties). Then we split the time of the trial, into time intervals like this: $\left[t(1), t(2)\right), \ldots, [t(r-1), t(r)), [t(r), t_{end})$ where t_{end} is the end of the trial. For each interval we need an estimation of the probability of a subject to not experience event, during that interval. We denote: n_i is the number at risk at $time = t(i)$. Also d_i is the number of the subjects that have experienced the event until $time = t(i)$. Then the Kaplan – Meier estimator is:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right). \tag{2.3}$$

$Rich\,et\,al.$ (2010) explained that due to the fact that the Kaplan-Meier estimator uses the observed events with in a non-parametric way, we cannot use it to extrapolate beyond the latest event time. It is also clear from the computation of the estimator that if we use it beyond the last event, we would be biased.

Another method for modeling and estimating the survival function is to use a parametric model for its form. We use known distributions (Exponential, Weibull, Gamma, Normal etc.) to model the survival distribution (equation 1.1) and then estimating the parameters with the method of the maximum likelihood. Now we will describe this method. Suppose that we have $t_1, t_2, \dots t_n$ observations from a certain distribution with p.d.f. $f(t;\lambda)$ where λ is an unknown parameter or vector of parameters. Then the general form of the likelihood function is:

$$L(\lambda; t_1, t_2, \dots, t_n) = \prod_{i=1}^{n} f(t_i;\lambda). \tag{2.4}$$

And then we maximize the likelihood function with respect to λ. The value that maximizes the function is the maximum likelihood estimator of the parameter. However, in the context of survival analysis there are censored observations. So we have to modify the general form of the likelihood function, in order to also take into account those observations. We will use the method described in Section 2.6 (page 21) of *Moore* (2016). For the censored observations we will use the survival function instead of the p.d.f., because from this observation we only know that the patient survived until a certain point. So the new form is:

$$L(\lambda; t_1, t_2, \dots, t_n) = \prod_{i=1}^{n} f(t_i;\lambda)^{\delta_i} S(t_i;\lambda)^{1-\delta_i}. \tag{2.5}$$

Note that if an observation is uncensored then $\delta_i = 1$ so $1 - \delta_i = 0$ and its contribution to the likelihood function is $f(t_i;\lambda)$. If an observation is censored then $\delta_i = 0$ and its contribution to the likelihood function is $S(t_i;\lambda)$. By maximizing the likelihood function with respect to λ we have an estimation for λ. Then, with this estimation and the equations (2.1) and (2.2), we have an estimation of the survival function.

There is also an estimation based on the Cox model that we are going to discuss later on this chapter.

We will also present two descriptive statistics based on the survival function, the mean and the median survival time. The mean survival time is the expected survival time of a patient and it is given by the following formula:

$$\mu = \int_0^\infty t f(t)\, dt. \tag{2.6}$$

From the equation (2.2) we have:

$$\frac{dF(t)}{dt} = f(t) \overset{(1.2)}{\Longleftrightarrow}$$

$$\frac{d(1 - S(t))}{dt} = f(t) \Leftrightarrow$$

$$f(t) = \frac{-dS(t)}{dt}. \tag{2.7}$$

Also, from $Moore$ (2016) Section 2.4 (page 15) we have that:

$$\lim_{t\to\infty}(tS(t)) = 0. \tag{2.8}$$

So if we combine the equations (2.6), (2.7) and (2.8) and we do the integration of (2.6) we have the following formula for the mean survival time:

$$\mu = \int_0^\infty S(t)\, dt. \tag{2.9}$$

The median survival time, is the time t such that $S(t) = 1/2$. However the survival function might not be continuous at ½. Then the median is the smallest t such that $S(t) \leq 1/2$. A non parametric estimation of the median survival time is the time that the Kaplan Meier curve crosses the 50% line. If the curve does not reach that point, then the estimator cannot be computed.

2.4 Hazard Function

The hazard function is the probability that, given that a subject has not experienced an event until time t, it will experience the event in the next very small time interval, divided by the length of that time interval. Because most of the times the event is death or failure, it is also called instantaneous failure rate. So the hazard function is:

$$h(t) = \lim_{\delta\to 0} \frac{P(t<T<t+\delta|T>t)}{\delta}. \tag{2.10}$$

$$\text{Therefore } h(t) = \frac{f(t)}{S(t)}. \tag{2.11}$$

And from the previous equation we have an expression for the p.d.f:

$$f(t) = h(t)S(t). \tag{2.12}$$

And for continuous distributions we have:

$$\frac{dh(t)}{dt} = -\frac{dS(t)}{dt}\frac{1}{s(t)} = -\frac{d\log S(t)}{dt} \overset{S(0)=1}{\Longrightarrow} S(t) = \exp\left(-\int_0^t h(u)du\right). \tag{2.13}$$

We define the function:

$$H(t) = \int_0^t h(u)du. \tag{2.14}$$

as the cumulative hazard function and it is the area under the hazard function up to time t. From the equations (2.12) and (2.13) we can derive the formula:

$$S(t) = \exp\bigl(-H(t)\bigr). \tag{2.15}$$

So the hazard function and the survival function are two ways to express a survival distribution that are connected with the formula (2.10). As we will see later, those two functions can be used for comparing different populations (e.g. populations that get different treatments for the same disease).

Now we will show the form of the hazard and survival function from 2 widely used distribution, the Exponential and the Weibull. The Exponential distribution is a special case of the Weibull distribution because it is a Weibull distribution with shape parameter 1. The formulas are taken from Section 2.4. (pages: 15-16) from *Moore* (2016).

The **Exponential** distribution with parameter λ has hazard function:

$$h(t) = \lambda. \tag{2.16}$$

In Figure 2 we visualize the hazard function of the Exponential distribution with parameter $\lambda = 1$ against time.

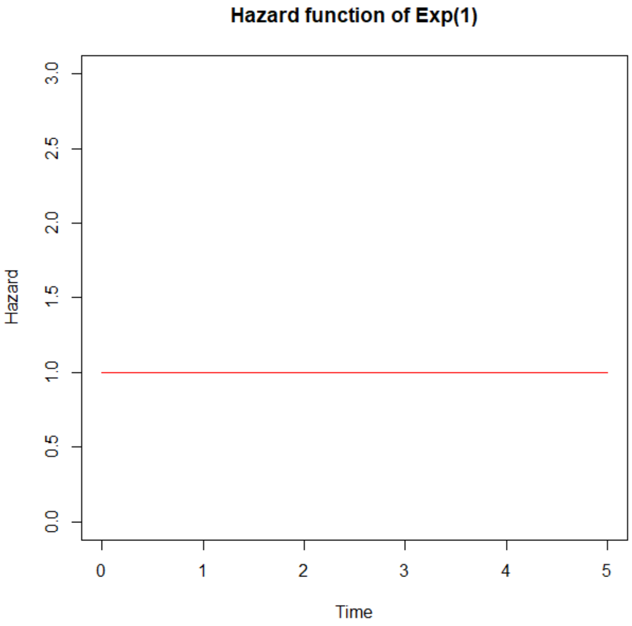


Figure 2: Hazard function of Exponential(1) distribution

Note that the hazard function of the exponential is constant. That makes the procedure simpler but it is also an assumption that rarely holds.

The cumulative hazard function from equation (2.14) is:

$$H(t) = \int_0^t h(u)du = \lambda t. \tag{2.17}$$

And from the equation (2.15) we get the survival function:

$$S(t) = \exp(-\lambda t). \tag{2.18}$$

We now visualize the Exponential distribution with $\lambda = 1$ against time.

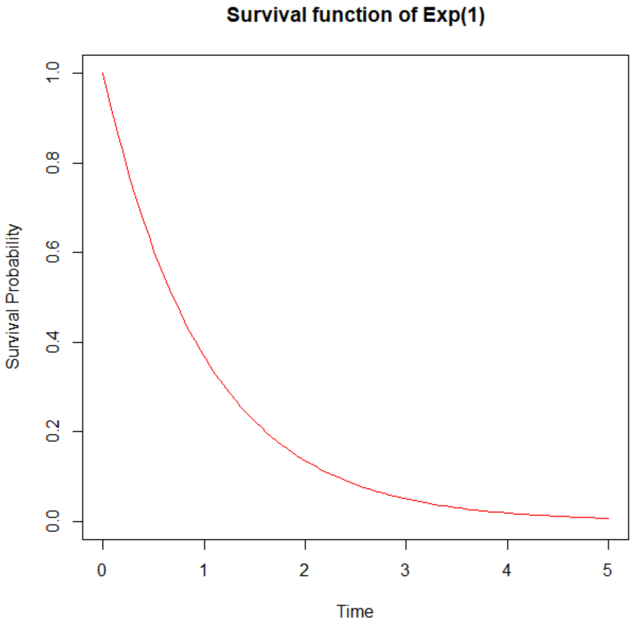


Figure 3: Survival function of Exponential (1)

From equation (2.12) we get a formula for the p.d.f.:

$$f(t; \lambda) = \lambda \exp(-\lambda t). \tag{2.19}$$

The **Weibull** distribution with shape parameter a and scale parameter λ has hazard function:

$$h(t) = a\lambda^{-a}t^{a-1}. \tag{2.20}$$

Note that for $a > 1$ the hazard function is monotone increasing and for $a < 1$ monotone decreasing. So the Weibull distribution gives us flexibility in modeling the hazard function. Because with the same distribution we can model both increasing and decreasing hazard functions, while the Exponential distribution ($a = 1$) only models constant hazard functions.

Now we will visualize (Figure 4) two different hazard functions in order to show the flexibility that the parameter a gives. The first will be the hazard function of Weibull (0.9, 1) and the second will be the hazard function of Weibull (1.1, 1). The first one has $a < 1$ and the second one $a > 1$.

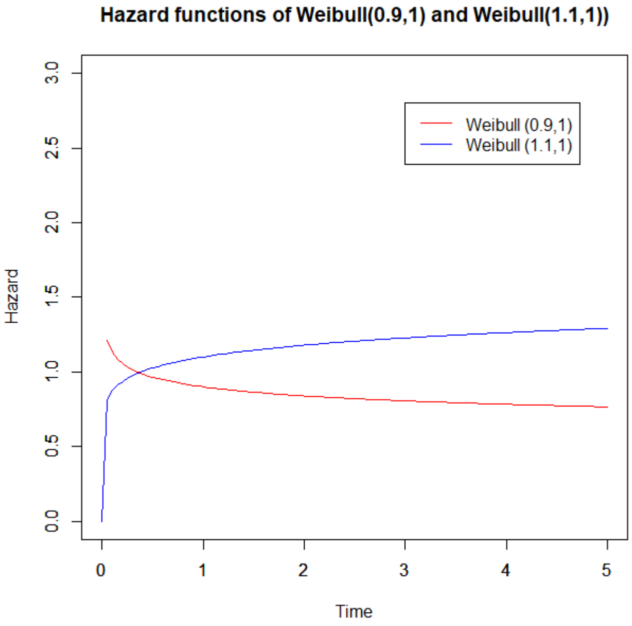


Figure 4: Hazard functions of Weibull (0.9,1) and Weibull (1.1,1)

The flexibility that the shape parameter gives us allows us to model a wide variety of situations with the Weibull distribution. We can have both increasing and decreasing hazards and this makes the model quite interesting for applications. For shape parameter equal to 1 we derive constant hazard as the distribution coincides with the exponential.

The cumulative hazard function from the equation (2.14) is:

$$H(t) = \int_0^t h(u)du = (\lambda t)^a. \qquad (2.21)$$

From the equation (2.15) we get the survival function:

$$S(t) = \exp(-(\lambda t)^a). \qquad (2.22)$$

Now we will visualize the survival functions of the Weibull (0.9,1) and Weibull(1.1,1) (Figure 5).

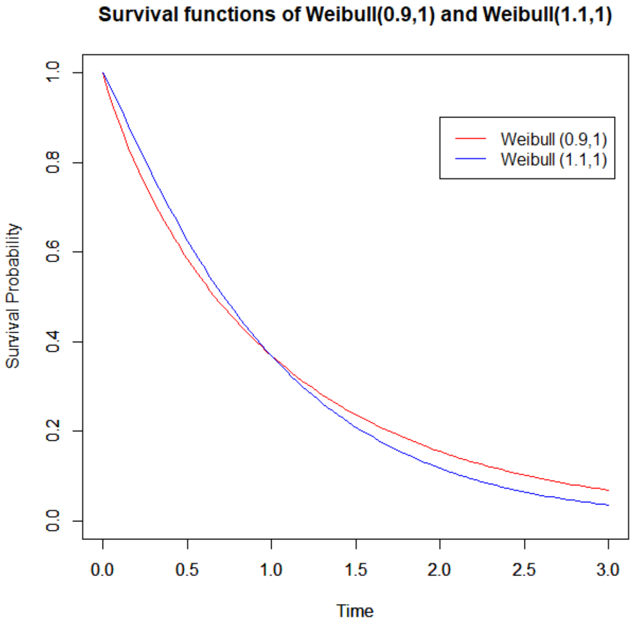


Figure 5: Survival functions of Weibull (0.9,1) and Weibull (1.1,1)

Note that at the beginning the red curve is under the blue which means that the population that has this survival function is dying “quicker” than the other population. However, after a while we have the exactly opposite situation. That happens because their hazard functions are crossing as we saw at Figure 5.

And from the equation (2.16) we get the formula of the p.d.f.:

$$f(t; a, \lambda) = a\lambda^a t^{a-1} \exp(-(\lambda t)^a). \qquad (2.23)$$

2.5 Proportional Hazards

We will now describe a very common assumption that is used for comparing different populations in the survival analysis. The idea is to assume, that the hazard function of each population, is proportional to the hazard functions of the other populations. So for the simple case of 2 populations, the assumption indicates that the rate of the instantaneous failure rates of the 2 populations is constant. This is a very strong assumption and later we will see how to test it. However it is very commonly used because of its simplicity. We will now express it more formal, for the general case of I populations. Let $h_i(t)$ be the hazard function of the i_{th} population. Then we assume that there is a common "baseline" hazard function $h_o(t)$, that all other functions are multiplications of it. We say that:

$$h_i(t) = h_0(t)\psi_i. \tag{2.24}$$

Where ψ_i can be a constant or a function. In most cases, we use $\psi_i = \exp(z_i b)$ in order to use covariates to explain the hazard function. In $\exp(z_i b)$, z_i is the value that the covariate takes for the i_{th} observation and b is the covariate's coefficient. It is clear that this assumption is a semi-parametric form, because we do not make parametric assumptions for the baseline hazard, so we do not assume its form. From the equation (2.24) we will derive some results as described from $Hosmer\ et. al.$ (2008) at Section 3.2 (page 71). For the cumulative hazard we have:

$$H_i(t) = \psi_i \int_0^t h_0(u)\, du = \psi_i H_0(t). \tag{2.25}$$

And we will denote $H_0(t)$ the baseline cumulative hazard function. From the equations (2.24) and (2.15) we have this formula for the survival function:

$$S_i(t) = \exp(-\psi_i H_0(t)) =$$

$$\exp(-H_0(t))^{\psi_i} =$$

$$S_0(t)^{\psi_i}. \tag{2.26}$$

And we will denote $S_0(t)$ as baseline survival function.

Now we need to estimate b in ψ. Because we have not assumed a parametric form for the baseline hazard we cannot use the classical likelihood theory. We will use the partial likelihood theory as presented by Cox (1972).

We will present the partial likelihood for the simple case where we only have two populations (e.g. two different drugs for the same disease) and the covariate is just the categorical that has $z_i = 1$ for i in population 1 and $z_i = 0$ for i in population 2. We will only use the failure times, and we will denote the j_{th} failure time as j and the time that happened t_j. So the hazard function of the i_{th} subject at that time is $h_i(t_j) = h_o(t_j)\exp(z_i b)$. Just before the first failure time t_1, we denote the set of the subjects at risk as R_1. So at t_1 a subject fails. The probability of that subject be the subject I is:

$$p_1 = \frac{h_i(t_1)}{\sum_{k\epsilon R_1} h_k(t_1)} = \frac{h_0(t_1)\psi_i}{\sum_{k\epsilon R_1} h_0(t_1)\psi_k}. \tag{2.27}$$

So we define like this all the probabilities for $j = 1 \ldots r$ where r is the total number of events. The partial likelihood is:

$$L(\psi) = p_1 p_2 \ldots . p_r. \tag{2.28}$$

We can see that the partial likelihood does not use the censored observations. Also, from the definition of p_j, we can see that the baseline hazard does not affect them because it is also at the numerator and the denominator. Therefore, we get the estimation of ψ by maximizes the partial likelihood with respect to b.

After the estimation of b we also need a test for it, and confidence intervals. The most commonly used test is the Wald test, and we can also construct a test statistic in order to give confidence intervals for the estimated covariates. We will give the description that is given by $Moore$ (2016) (page 60). The form of the test statistic will be:

$$\frac{\hat{b}}{s.e.(\hat{b})}. \tag{2.29}$$

Where $\hat{b}$ is the value that maximizes $L(b)$ from the equation (2.28) and $s.e.(\hat{b})$ is the standard error of the estimation. To compute the standard error we have to use the "observed information" which minus the second derivative of $\log(L(b))$. We will denote it:

$$I(b) = -\frac{d^2(\log(L(b)))}{db^2}. \tag{2.30}$$

And we have that:

$$var(\hat{b}) \cong \frac{1}{I(b)} \text{ and } s.e.(\hat{b}) \cong \frac{1}{\sqrt{I(b)}}. \tag{2.31}$$

Now we construct a normalized test statistic $Z_w = \frac{\hat{b}}{s.e.(\hat{b})}$ for testing the hypothesis $H_0: b = 0 \; vs \; H_1: b \neq 0$ and we reject the null hypothesis if $|Z_w| > Z_{a/2}$. The $1-a$ confidence interval is $[b\hat{} - Z(a/2)\, s.e.(b\hat{}), b\hat{} + Z\_(a/2)\, s.e.(b\hat{})]$. We can also do the test with the ${Z_w}^2$ because it follows a Chi-square distribution with 1 degree of freedom. So we reject the null hypothesis if ${Z_w}^2 > X_{a,1}^2$.

From the proportional hazards model, we can also estimate the survival function, if we can estimate the baseline survival function. *Moore* (2016) at Section 5.5 (page 64) proposes an estimator for the baseline hazard function:

$$\widehat{h_0}(t_i) = \frac{d_i}{\sum_{j \; in \; R_j} \exp(\widehat{\psi_J})}. \tag{2.32}$$

Where d_i is the number of events up until t_i and R_j is the set of all the patients that are at risk at the time point t_i. From this estimation we can have an estimation for the cumulative hazard function at time t by adding the estimated $h_0(t_i)$ for $t_i < t$. Then with the last estimation we can estimate the baseline survival function. And by combining this estimation with the estimation of ψ_i in the equation (2.28), we have an estimation of the survival function given by the proportional hazards model.

2.6 Log-Rank test

Until now we have described different aspects of expressing the survival profile of different populations. We would also like to have a method to compare different populations (e.g. different treatments for the same disease). More specifically, for 2 populations we would like to test the hypothesis $H_0: S_1(t) = S_2(t)$ against $H_1: S_1(t) \neq S_2(t)$ where $S_i(t)$ is the survival distribution of the i_{th} population. However, in survival analysis there is a huge amount of alternatives to this hypothesis because the survival functions can have a wide variety of functional form. Therefore, we will restrict those alternatives to the ones that can be expressed as:

$$H_A: S_1(t) \neq S_2(t)^{\psi}. \tag{2.29}$$

Or into the equivalent hazard form:

$$H_A: h_1(t) \neq \psi h_2(t). \tag{2.30}$$

So it is equivalent to the proportional hazard assumption. Now the hypothesis testing can be expressed as:

$$H_0: \psi = 1 \; vs \; H_A: \psi < 1. \tag{2.31}$$

So if the alternative hypothesis is true, then the population in group 2 will have uniformly longer survival times than those in group 1. We can do this hypothesis testing with the Log-Rank test. As we mentioned before, the proportionality assumption is very strict. So if we use the Log-Rank test we should firstly test if this assumption holds. A test for that purpose will be presented later in this chapter. We will now describe it in a formal way.

Suppose that we have 2 populations and let t_i be the i_{th} failure time from both populations. Then at $time = t_i$ we denote:

$\mathrm{d}_{1\mathrm{i}}$ = number of failures from the first population

$\mathrm{d}_{2\mathrm{i}}$ = number of failures from the second population

n_{1i} = number at risk from the first population

n_{2i} = number at risk from the second population

$n_i = n_{1i}+n_{2i}$

$d_i = d_{1i}+d_{2i}$

At every failure time we can construct a table like this:

	Population 1	Population 2	Total
Failure	d_{1i}	d_{2i}	d_i
Non Failure	$n_{1i}-d_{1i}$	$n_{2i}-d_{2i}$	n_i-d_i
Total	n_{1i}	n_{2i}	n_i

We want to check the independence of the failure numbers in the two groups. If they are independent then there is no difference in the survival functions of the two populations. If they are indeed independent, and for given d_i,n_i,n_{2i}, n_{1i}, we have that d_{1i} follows the hypergeometric distribution. Thus:

$$p(d_{1i}|d_i,n_i,n_{2i},n_{1i}) = \frac{\binom{n_{1i}}{d_{1i}}\binom{n_{2i}}{d_{2i}}}{\binom{n_i}{d_i}}. \tag{2.32}$$

Therefore, it is very easy to compute the mean and variance of d_{1i}. Note that d_{1i} takes values in $[0, N_1]$ where N_1 is the number at risk in population 1 at the beginning of the trial. We have:

$$E(d_{1i}) = \frac{n_{1i}d_i}{n_i} \tag{2.33}$$

$$Var(d_{1i}) = v_{1i} = \frac{n_{1i}n_{2i}d_i(n_i-d_i)}{n_i{}^2(n_i-1)}. \tag{2.34}$$

Now we denote the mean d_{1i} with e_{1i} from the word expected and its variance v_{1i}. Then we sum over all the tables the difference between the expected and the observed values:

$$U_0 = \sum(d_{1i}-e_{1i}). \tag{2.35}$$

With variance given by:

$$Var(U_0) = \sum v_{1i} = V_0. \tag{2.36}$$

So now we know that asymptotically $\frac{U_0}{\sqrt{V_0}} \sim N(0,1)$ and $\frac{{U_0}^2}{V_0} \sim X_1^2$. And with the last statistic we do the hypothesis testing.

$Harrington\ and\ Fleming$ (1982) proposed a generalization of the log Rank test. The generalization is based on the use of weights in the equations (2.35) and (2.36). The idea is that in the summation of U_0 we might want to add more weight to the patients that experience the event early in the trial and less to those that experienced the event later. The weight function that they proposed is:

$$w_i = \hat{S}(t_i)^{\rho}. \tag{2.37}$$

And the new form of the statistic is:

$$U_0 = \sum w_i(d_{1i} - e_{1i}). \tag{2.38}$$

With variance given by:

$$Var(U_0) = \sum {w_i}^2 v_{1i}. \tag{2.39}$$

Where $\hat{S}(t_i)$ is the Kaplan – Meier estimate of the survival function. The weight function is a function of the estimation of the survival function. Suppose we have a patient that died at the beginning of the trial. Then just before his death, his survival function (the probability that his life will exceed this point) had a high value. In contrast, the survival function of a patient that died later, at the moment of his death was lower. So the weight function achieves to make the early deaths more "important" (depends on the value of ρ how much) in the process of detecting a difference in the survival curves. This family of weighted Log Rank tests is called Fleming-Harrington $G(\rho)$ tests. Note that for $\rho = 0$ we have the Log Rank test.

The value ρ determines how important will be the time of death. For example, with values close to 0, the earlier deaths are a lot more important and with values close to 1 the weight function is more conservative. The reason behind this is that we might be more interested, for clinical reasons on the early events. For example, if we are studying a population of elderly people, we might be more interested in the early stages of the trial. That is because at later stages of the trial and due to their age, the patients are more likely to die, independently from the illness that we are studying.

Now we will conduct a simulation study, in order to compare the power that we have from the Log Rank test and the Fleming-Harrington $G(\rho)$ test for two different values of ρ, one small and one big. We choose a small and a big value in order to investigate the two opposite cases of the test and by that we will understand it better. The two values will be 0.2 and 0.8. For the simulation, at each repetition, we will simulate two survival populations with proportional hazards, and we will choose a censoring distribution for achieving 22% censoring of the total sample. The censoring distribution is chosen by simulating the two survival populations with a censoring distribution and we measure the percentage of censored observation. Finally we keep the censoring distribution that has mean percentage close to 22%. We will test the hypothesis $H_0: S_1(t) = S_2(t)$ against any alternative at significance level $a = 0.05$ and we will get the Monte Carlo estimation of the power for each test. That will happen by simulating both populations with the censoring distribution and then test the hypothesis. The percentage of the repetitions that the null hypothesis is rejected will be the estimated power. The repetitions will be 1000. The details are presented on the next table. The sample size will be 90 patients at each population.

Table 1: Details of the simulation.

Population	Distribution
Population 1	Exponential (1.1)
Population 2	Exponential (1.8)
Censoring	Weibull (2,3.3)

Now we will visualize the hazard and the survival functions against time, in the following Figures. 6 and 7 respectively

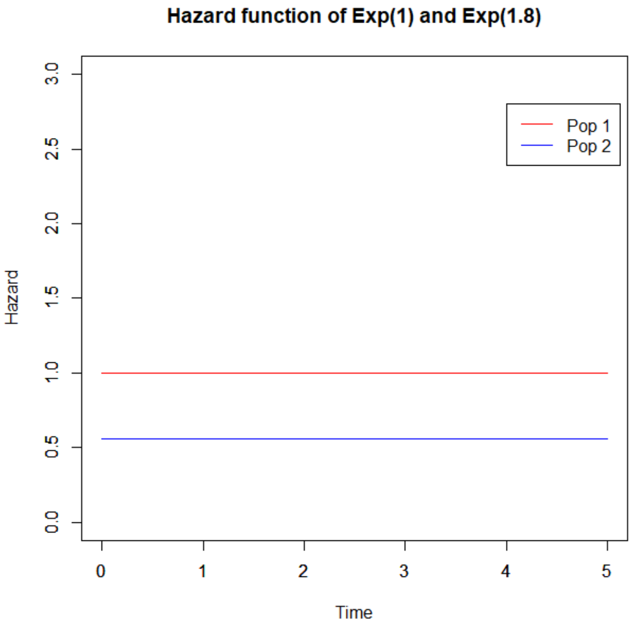


Figure 6: Hazard function of Exponential(1) and Exponential(1,8)

Note that the two hazard functions are constant, because the exponential function has constant hazard function, as we mentioned before.

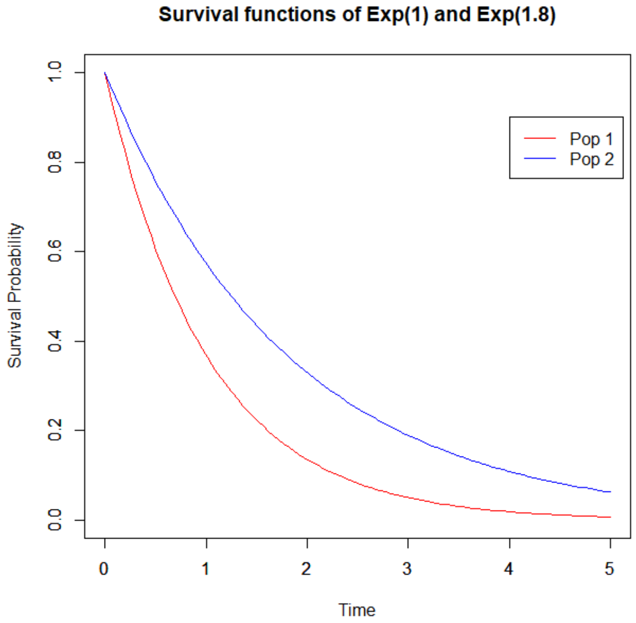


Figure 7: Survival functions of Exponential(1) and Exponential(1.8)

The results are presented in the next table.

Table 2: Results of the simulation.

TEST	POWER
Log Rank	89.7%
$G(\rho), \rho = 0.2$	90.1%
$G(\rho), \rho = 0.8$	85.4%

From the results we can see that the Log Rank test is almost identical with the Fleming-Harrington test for $\rho = 0.2$, while it is better than it for $\rho = 0.8$. It is known that under proportional hazards the Log Rank test is the optimal choice. In chapter 4 we will compare the two tests under cases of non-proportional hazards as well.

2.7 Test for the proportionality assumption

The proportionality assumption is a very convenient tool for modeling and testing differences between survival curves of different populations. In the previous sections of this chapter we discussed the basics of the Cox model and the Log-Rank test. However, despite the simplicity that it provides, it is a very strict assumption that rarely holds. So we will present in this section a test for this assumption.

First of all we will denote the Schoenfeld residuals that were proposed by $Schoenfeld$ (1982) for the Cox model. Let $T_1, T_2, \dots, T_n$ be the observed survival times and $\delta_1, \delta_2, \dots, \delta_n$ their censoring indicators. Let $z_1, z_2, \dots, z_n$ be vectors of fixed covariates, $b \, and \, \hat{b}$ the unknown coefficients and their usual estimations and R_i be the risk set at time $t_{(i)}$. We will denote the i_{th} observed time as $t_{(i)}$ and $Z_{(i)}, R_{(i)}$ the corresponding covariate vector and risk set. The Schoenfeld residuals are:

$$\hat{r}_{(i)} = z_i - \frac{\sum_{j \, inR_{(i)}} z_j \exp(\hat{b}^T z_j)}{\sum_{j \, inR_{(i)}} \exp(\hat{b}^T z_j)}. \tag{2.40}$$

$Grambsch \, and \, Therneau$ (1994) transformed the Schoenfeld residuals and proposed a test based on them for the proportionality assumption. They proposed the scaled Schoenfeld residuals which are:

$$\hat{r}_{(i)}{}^* = \frac{\hat{r}_{(i)}}{\bar{V}}. \tag{2.41}$$

$$\text{Where } \bar{V} = \frac{\varphi(\hat{b})}{d}. \tag{2.42}$$

Where d is the total number of uncensored events and $\varphi(\hat{b})$ is minus the second derivative of the log likelihood of the Cox model.

For the test statistic let $g_i \, and \, \bar{g}$ be the time scale and the average time scale respectively (either linear or logarithmic) and I_k the information matrix elements for covariate K. Then the statistic for covariate K is:

$$T_K = \frac{\left(\sum(g_i - \bar{g})\hat{r}_{(i)}{}^{*}\right)^2}{dI_k\left(\sum(g_i - \bar{g})\right)^2}. \tag{2.43}$$

And the statistic in the covariate specific form follows X_1^2 distribution. Extreme values of this test is an indication against the proportionality assumption.

To present this test we will give an example with simulated data. The simulation will be from the same set up as in table 1 where the hazard functions are proportional. In the next Figure we will show the Kaplan Meier estimators from the 2 simulated populations.

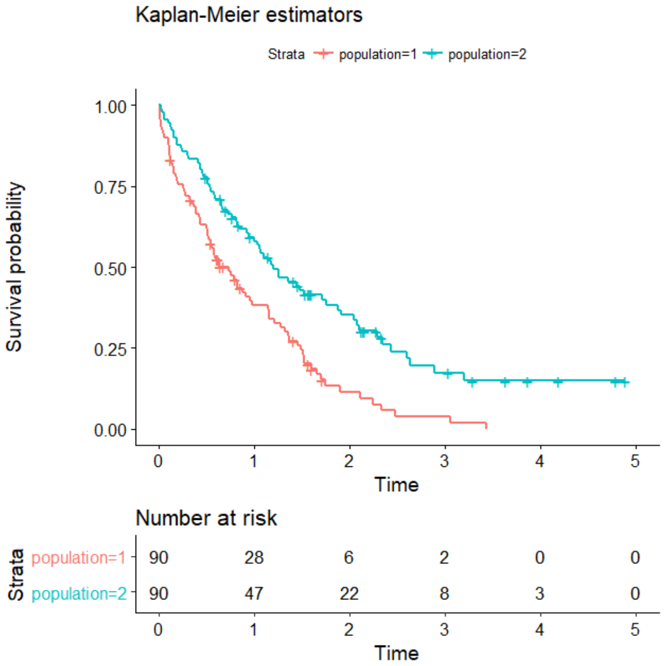


Figure 8: Kaplan Meier estimators for simulated data from table 1

The statistic T takes the value 0.177 and the p-value is 0.674. So we do not reject the proportionality assumption which is correct.

For different tests one can see the work of Grant et al (2014). In this paper, also the power of the tests have been examined through simulations showing that the performance of the test above is not very good especially for some particular scenarios.

We can see this also at the simulation study in section 5.4 (Table 16). This implies also the need for further improvement in the methodology presented in this thesis.

CHAPTER 3: Clinical Trials

3.1 Introduction

Clinical trials are studies, where human subjects are assigned and test the risks and the benefits of new medical therapies. National Institutes of Health of USA gives the following definition for clinical trials: “A research study in which one or more human subjects are prospectively assigned to one or more interventions (which may include placebo or other control) to evaluate the effects of those interventions on health-related biomedical or behavioral outcomes”. Since there are human participants, there are very strict regulations for conducting a clinical trial. It is obligatory, that a protocol is made at the beginning of the trial that describes the targets and the whole procedure, and the researchers must strictly follow it. Because of those regulations, designing a clinical trial is a very complicated task because there are many parameters that the researches have to take into account. We will now describe in more detail those regulations.

First of all, the researchers have to define the clinical outcome that the want to study. For example, it might be the survival times of the participants that take a certain drug. The new medical methods are compared to others that already exist. Therefore the researches should also define a summary measure that will compare the methods (e.g. difference of median survival times, hazard ratio, difference in Restricted Mean Survival Times or other biological measures) and also a targeted value of that summary measure that they want to detect (e.g. hazard ratio = 1.25). Since the outcome will be derived with statistical testing, the researchers should specify the significance level and the power of the test. Meaning that when they compare new interventions to standard ones, they should specify the probability they will falsely claim superiority and the probability that they will not detect a possible existing superiority. Also, they must specify when the trial is going to end, and combining this with the previous, to estimate the number of participants or number of events that the trial needs to meet up the design specifications.

There are also the regulations about the participants. Since there are risks in the testing of new medical therapies in clinical trials, they must not exceed the possible benefits from the new medical therapy. The experimental design must not expose subjects to unnecessary dangers and the procedure should not be prolonged beyond a certain point where the goals have been achieved. However, the number of patients and the duration of the trial, must be large enough in order to test sufficiently the effect of the therapy and give reliable and unbiased results. The gathered medical data, should be used in a way that the privacy of the participants is secured. Also, the participants that take part in a clinical trial must be fully informed about the targets of the trial and also the possible benefits and risks. They are also allowed to stop their participation in the trial at any point.

Conducting a clinical trial is the most efficient way to test if a new treatment is more effective than the existing one, and also to test adverse effects that might have. For a drug to become available to the public, it has to be tested with a clinical trial and to have sufficient evidence that it is more effective than the current treatments. Then, based on that evidence, it will get approval from a public organization responsible for drugs (e.g. FDA).

3.2 Types of clinical trials

When a new treatment is designed at the lab, then it has to be tested on humans. There are many aspects than need to be specified, such as the dose of the new drug or the way that the patients will receive it. Of course, as we discussed before, the new method needs to be more efficient than the current one and to not have adverse effects. Now we will present the phases that test a new treatment as described by *Cook and DeMets* (2008) in Section 3 (pages 75-90).

1) **Preclinical:** The researchers identify a risk factor related to a certain disease. Then they try to make an intervention that modifies the risk factor. In those studies testing often includes animals. If they have sufficient evidence that a new intervention for a specific disease is promising, then they continue with clinical trials.
2) **Phase 1:** *Meinert* (1996) gave the following definition: "Broadly, a trial involving the first applications of a new treatment to human beings and conducted to generate preliminary information on safety". Generally, at this phase, we want to test if the treatment is safe to be given at humans and also in what dose and way. We test the toxicity of each dose by examining the adverse effects that has on patients. Also we examine what is the best way to apply the new intervention (e.g. pills or injection). The sample sizes are often between 20 and 50.
3) **Phase 2:** At this phase, the treatment is tested for its effectiveness and possible adverse events. Generally, phase 2 trials aim to detect the possibility that the new intervention can be superior to the standard one. They also examine the adverse effects like in the previous phase. The main objective of those trials is to examine if a new trial, larger in sample size and duration should start. So typically phase two trials have short duration and sample size below 60 patients.
4) **Phase 3:** At this phase, the treatment is compared to the current treatments again in a more detailed way. The trial is larger in sample size and duration and the outcomes are more accurate. The new treatment needs to be more effective in order to be approved.
5) **Phase 4:** Those trials are conducted after the approval of a new intervention (after phase 3) and intend to gather additional information about its efficacy or safety.

Now we will discuss the different designs that exist in clinical trials. When we want to compare a new treatment with the existing one (we will call it control group), we have to get data from populations that get those treatments. So we have to select patients for both treatments, with respect to time and budget limits. The different designs are:

1) **Historical:** We give the new treatment to patients and we gather the data. However, the data that we need from the control group are gathered from previous trials, that the control treatment was used. However this type of trial is often biased. That happens because, most of the times the two trials are not identical, meaning that they might have different definition, diagnostic criteria or duration. So the outcomes that are compared are not produced from identical procedures.
2) **Concurrent:** In this design we give the new treatment to patients and then gather the data. At the same time, we gather data from patients that do not get the treatment. However, the two groups are not in the same clinic or site. That eliminates the sources of bias that the historical trials cause, because the clinicians can design the trial from the start. It also helps with the sample size, because each clinic needs to gather people only for one group and not for all of them. However there is the problem of selection bias. Each clinic specializes in different areas and level of treatment. For example tertiary referral centers have different expertise than primary care facilities, so they choose patients accordingly. That makes the groups to contain different mixtures of patients.
3) **Randomized:** In this design, we enlist patients to the trial, and then with a randomized process we assign them to each treatment group. The randomization process deals with the bias of selection. Each group has the same mixture of patients and can be compared to each other without any source of bias. Because of that, randomized clinical trials are the safest option for a design.

A clinical trial design should be very strict to its target and its procedure, as we discussed previously. However, this can escalate the cost and the complexity of the trial to a point that it may even be cancelled by the pharmaceutical company that conducts it. Furthermore, a clinical design must stop or be modified if it has certain outcomes before the planned ending. For example, if the examined treatment turns out to be harmful and the adverse events that causes outweight the possible benefits. Or at some point, the superiority or the inferiority of the treatment that we are studing is established. At cases like the previous, the trial must be terminated. For solving all of these problems, we need to add flexibility to the designs.

Mahajan and Gupta (2010) described the concept of **adaptive designs** to add flexibility, efficiency and speed to clinical trials. To achieve them we need to be able to modify the trial for keeping its cost and complexity at acceptable level and the participants safe. In particular, we need to be able to modify the trial procedure, or the statistical procedure or both during the conduct of the clinical trial. However, those modifications should not undermine the scientific validation of the trial. The modifications have to be pre-planned and be based on the analysis of the interim data from the study. That means that the modifications must be planned before the data examination, which means that they must be contained in the initial design. Now we will present some adaptive clinical designs that *Mahajan and Gupta* (2010) give.

1) **Adaptive randomization design:** With this design, we firstly make a randomization procedure as always, but we are allowed to alter by using unequal probability of assignment in the different groups. We may adapt the randomization with criteria based on the treatment or other covariates and the response of the treatment.
2) **Group sequential design:** In this design, the patients are enrolled in groups and we conduct interim analysis in each group sequentially, and we decide if the trial must be terminated or modified to prevent safety or efficacy issues. In trials with this design, the researchers should be careful with the initial significance level that they have set. Due to the multiple testing of the interim analysis, the significance level may be not the one that was set at the beginning.
3) **Sample size re-estimation design:** In this design, we re-estimate the sample size based on the interim analysis. The criteria can be related to the treatment effect or the conditional power. However, researchers should be careful not to start the trial with a very small sample size. An interim analysis based on a small sample, may produce statistically insignificant outcomes and do the re-estimation of the sample size inaccurate.

4) **Drop-the-loser design:** In this design we want to exclude from the trial the patients that got the inferior treatment, and continue with the rest of them. So we usually split the trial into two parts, and in the first we give different treatments at the groups. Then we observe which treatment has no effect, based on pre-specified criteria and we exclude the patients of this group. Then we continue with the rest. The Drop-the-loser design is very helpful to detect the minimum effective and the maximum tolerable dose of a new treatment.
5) **Biomarker adaptive design:** In this design, we decide whether we will modify the procedure or not based on interim analysis of various biomarkers that are related to the disease that we are interested in. However, analyzing biomarkers as a criterion for modification of the trial is different than building a predictive model that relates the biomarkers with the clinical outcomes.
6) **Adaptive treatment-switching design:** In this design we are allowed to change the treatment group of a patient if there are safety or efficacy issues. However, this adaptation makes the estimation of the survival rate very difficult and makes adaptation of the sample size necessary.
7) **Hypothesis adaptive design:** This design allows us to change the initial hypothesis (e.g. to make a superiority hypothesis a non-inferiority)
8) **Adaptive seamless phase II/III design:** In this design we conduct the phase III right after the phase II. We use the data collected in phase II additionally to the data produced by phase III, and we save time and making the procedure easier. However, there is the problem of multi testing and its impact on the predefined significance level. Also, in many cases the two phases have different end points which also makes those designs questionable.
9) **Multiple adaptive designs:** This is any combination of the previous adaptive designs. However, in the multiple adaptive design statistical inference can be very complex.

3.3 Sample size calculations

A very important aspect of designing a clinical trial, is the number of patients that are needed, in order to achieve a certain significance level and power, given a targeted summary measure. Although, when we have time to event data, we first need to calculate the number of events that we need for reasons that we will explain later. In most cases the summary measure is the hazard ratio. However, this implies a very strong assumption about the hazard functions of the populations of the trial. The assumption is that the hazard functions of the populations are proportional, thus the hazard ratio is constant over time. There is an underlying hazard that is common, and we do not know its functional form. This assumption allows us to use the Log-Rank test and that provides formulas for the calculation of the number of events as presented by $Schoenfeld$ (1983). We will give the proof from that $Moore$ (2016) gives in Section 11.4 (pages 165-166). Suppose that we will perform an one-sided test. The targeted hazard ratio is Δ_O (assumed to be constant) and we want to have significance level a and power β. We also denote the k percentile of the standard normal distribution as Z_κ. Also, let P_A be the proportion of the sample allocated to treatment A and P_B to treatment B. We also have the notation of section 2.6 for number of failures and number at risk at each population at each time $(\mathrm{d_{1i}, d_{2i}, n_{1i}, n_{2i}, n_i, d_i})$ and let t_i be the i_{th} failure time from both populations.

For the computations we will use the Log Rank statistic U_0 as defined in equation 2.35. Its variance for the i_{th} failure time is:

$$v_{1i} = var(d_{1i}) = \frac{n_{1i}n_{2i}d_i(n_i - d_i)}{{n_i}^2(n_i - 1)}. \tag{3.1}$$

Now we will assume that the number of deaths at each time is relatively small in comparison with the number at risk, and that $P_A \cong \frac{n_{1i}}{n_i}$ is constant over time, we have approximately that v_{1i} from equation 3.1 is:

$$v_{1i} \cong \frac{n_{1i}n_{2i}d_i}{{n_i}^2} \cong P_A P_B d_i. \tag{3.2}$$

So the approximation of the variance of the Log Rank statistic U_0 is:

$$V_0 \cong P_A P_B \sum_1^D d_i \cong P_A P_B d. \tag{3.3}$$

And since we know that $\frac{U_0}{\sqrt{V_0}} \sim N(0,1)$, with trivial computations we find from *Moore* (2016) in Section 11.4 (page 166) that the expected number of deaths d in order to detect the targeted hazard ratio Δ_O on a significance level a and power β is:

$$d = \frac{(z_b + z_{1-a})^2}{P_A P_B \log_e \Delta_O{}^2}. \tag{3.4}$$

However, the clinical trials cannot last until all the patients experience the event. Therefore, if we want to calculate the total number of patients that we need in order to achieve this number of events, we need to make more assumptions. We need to assume a parametric form of the survival curves in both groups. We usually assume that both the survival distributions follow exponential distributions or Weibull distributions with the same shape parameter. Now we will show the procedure for group A as described in section 11.2 (page 161) of *Moore* (2016). The number of patients that we need is dP_A. We split the total period of the clinical trial into two parts. The first part is the accrual period c and the second part is the follow up period f. An individual can be assigned for the trial at any point in the time interval $[0, c]$, and we assume that this time follows the uniform distribution. The probability of an individual that enters the trial at time t assuming that the survival distribution of its group follows an Exp(λ) is:

$$\pi(t) = 1 - S(c + f - t; \lambda). \tag{3.5}$$

Since t follows Uniform (0,c) we can have the average probability by the following integral:

$$\int_0^c \frac{1}{c} \Pr(\, death | enter\ at\ time = t) dt = \int_0^c \frac{1}{c} (\, 1 - S(c + f - t; \lambda)) dt. \tag{3.6}$$

We use the variable transformation $c + f - t = u$ and from the assumption for the survival distribution we have that $S(u; \lambda) = \exp(-\lambda u)$. Therefore the probability of death at group A is:

$$\pi_A = 1 - \frac{1}{c}\int_f^{c+f} \exp(-\lambda u)\, du. \tag{3.7}$$

And that is equal to:

$$\pi_A = 1 - \frac{1}{c\lambda}\left[\exp(-\lambda f) - \exp\big(-\lambda(c+f)\big)\right]. \tag{3.8}$$

If we assume for the survival distribution a $Weibull(k,m)$ (note that we also assume that the other survival distribution is also Weibull with the same k because we assume proportional hazards) then we have the following formula that can be computed numerically:

$$\pi_A = 1 - \frac{1}{c}\int_f^{c+f} \exp\left(-\log(2)\left(\frac{u}{m}\right)\right) du. \tag{3.9}$$

But in clinical trials we also have censored observations. Some participants will not stay in the trial until the end. Let C be the censoring rate and we assume that it is known. Then the patients that we need for the group A are:

$$n_A = \frac{P_A d}{\pi_A C}. \tag{3.10}$$

3.4 Sample size calculations with R package powerSurvEpi

The package $powerSurvEpi$ contains the function ssizeCT that calculates the sample size needed based on a pilot data set. By pilot data set we mean a data set taken from a previous clinical trial that was very similar to ours. A function like that can be very useful, because sometimes there are data available from very similar clinical trials. From those data sets, the function calculates the probability of dying in each group and computes the sample size needed for given significance level, power and targeted hazard ratio. It also estimates the censoring rate. The only assumption that is made is the proportionality of hazards. Therefore, we avoid assuming a specific parametric form for the two curves in order to compute the probability of death. Later, we will use this function in order to study the miscalculations on sample sizes when the proportionality assumption does not hold.

Now we will describe the way that the function estimates the probability of death in each group. The function uses a method that was proposed by $Freedman$ (1982) and follows the description in Section 14.12 (page 807) of $Rosner$ (2006). The inputs of the function are:

1) The pilot survival data set
2) The targeted hazard ratio
3) The significance level
4) The power
5) The ratio of the participants in the experimental group over the participants of the control group

Then we denote as n_E, n_c the number of the participants in the experimental and in the control group and d the total expected events. We also denote the ratio of the participants $\frac{n_E}{n_c} = k$. The function estimates the probabilities π_E, π_C which are the probability of someone dying in the experimental and in the control group. The targeted hazard ratio is denoted as RR. The test that the calculations are based on is two-sided. The outputs are given by the following formulas:

$$n_E = \frac{dk}{\pi_E k + \pi_C} \tag{3.11}$$

$$n_C = \frac{d}{\pi_E k + \pi_C}. \tag{3.12}$$

And the expected number of total events $d = \frac{1}{k}(\frac{kRR+1}{RR-1})^2(z_{1-\frac{a}{2}} + z_{1-b})^2$.

(3.13)

3.5 The impact of the proportionality assumption on the sample size calculation

The previous methods for calculating the sample size are based on the assumption of the proportional hazards. That is the reason for specifying a targeted hazard ratio. If the proportionality assumption does not hold, the hazard ratio is a function of time. In this section we will do simulation studies, in order to estimate the miscalculations that might occur, when the assumption is wrong. We will simulate survival data sets, from clinical trials that compare two drugs (Experimental-Control) with equal allocations to each group. The hazard functions from the two groups will not be proportional. The end of the trial will be 1.3 times the median of the experimental drug. Then we will estimate the sample size that we need, in order to achieve 80% power on a 10% significance level. For this calculation, for different sample sizes we will generate 1000 Monte Carlo clinical trials and we will estimate the power, until we end up with a sample size that gives us 80%. The estimation of power is given by simulating the two groups in each repetition with a censoring given by the censor distribution and then do the test. The percentage of the repetitions that the test correctly rejected the equal hazards assumption will be the estimated power. The test will be the Log Rank and will be performed by the function survdiff in the package *survival*. Then we will compare it with the estimated sample size that we would have, with a method that assumes proportional hazards.

In order to minimize the assumptions that we make we will use the ssizeCT function from package powerSurvEpi, and the probability of death and the censoring rate will be estimated from a pilot data set. In a real clinical trial where a pilot data set does not exist the clinicians must assume the survival distributions for the sample size. However we are interested only in the impact of the wrong assumption of proportional hazards. So for the investigation we would like to eliminate all the sources that cause wrong estimation of the sample size (e.g. assuming wrong parametric forms). For this reason, the pilot data set for both groups will be simulated from the same survival functions for each group, the same censoring distribution and the same duration. By doing that, we construct a scenario where we have a perfect pilot data set and the only problem that causes wrong estimation is the assumption of proportional hazards. So the estimated probabilities of death and censoring will be accurate, while in real clinical trials it is not possible to have such a compatible pilot data set. For the targeted hazard ratio that we need as input, we will use the real hazard ratio at two different times. The times will be the medians of the two survival functions and we will have two estimated sample sizes. Also the proportionality assumption will be violated in three different ways:

1) Early diverging hazards
2) Late diverging hazards
3) Crossing hazards

For the first case (early diverging hazards), the simulation design will be:

Table 3: Details of the simulation study design

Drug	Distribution	Censoring percentage	Median
Control	Weibull (0.9,0.9)	14%	0.59
Experimental	Weibull (1.2,1.5)	23%	1.10
Censoring	Weibull (2,2.9)	-	-

Now we will visualize the hazard functions against time, and we will add the two median times in the next Figure. We can see that the hazards have greater distance at the beginning and then they converge. However, at the time of the crossing, most of the participants have already experienced the event. So the crossing that occurs in this case is of no importance.

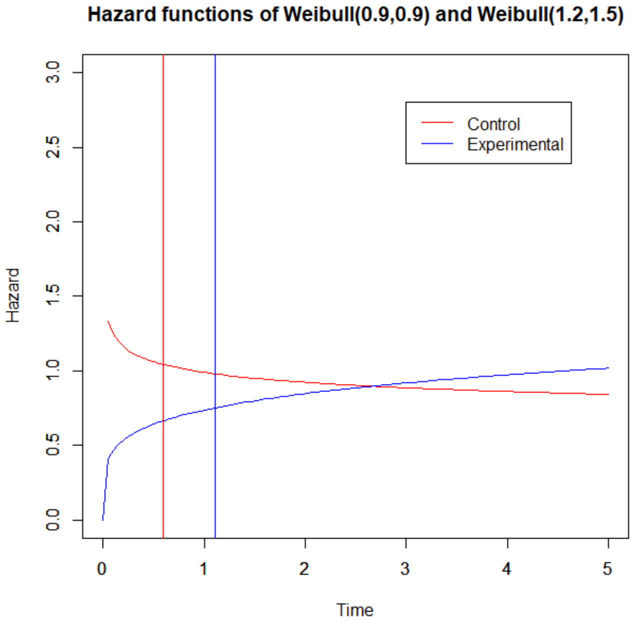


Figure 9: Hazard functions (the curves) and medians (the vertical lines) of Weibull (0.9,0.9) and Weibull (1.2,1.5)

The expected Kaplan Meier estimators are expected to be like the theoretical survival curves in (Figure 10). We also add the two medians by drawing a horizontal line that crosses the y axis at 0.5

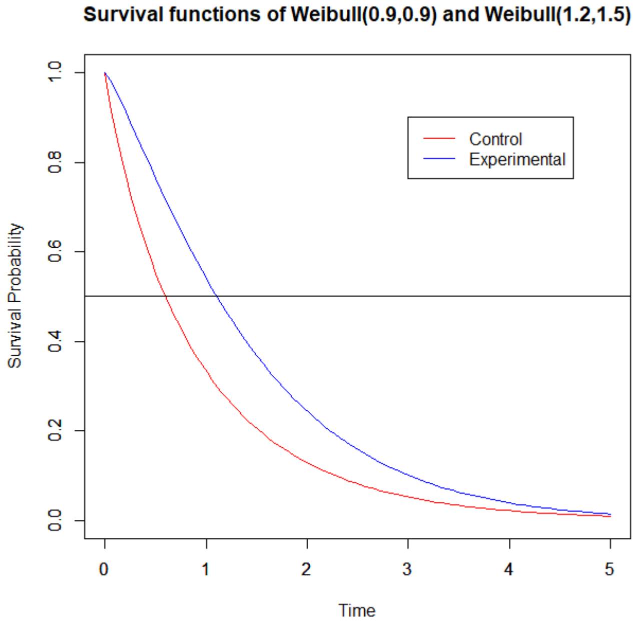


Figure 10: Survival functions (the curves) and medians (the points that the horizontal line crosses each curve) of Weibull (0.9,0.9) and Weibull (1.2,1.5)

As we described before, we will find the sample size that achieves power 80% through simulations. The one that achieves this power is 114. The estimated results from the ssizeCT function for the two different hazard ratios are:

Table 4: Results from the powerSurvEpi package

Time	Hazard Ratio	Estimated Sample Size from ssizeCT
Control Median (0.59)	0.63	182
Experimental Median (1.10)	0.76	482

It is clear that the violation of the proportionality assumption lead to huge overestimations of the sample size that we need. The actual size that we need is 114 and the estimations were 182 and 482. In clinical trials, where the participants might be hard to find, those miscalculations can be very harmful.

For the second case (late diverging hazards), the simulation design will be:

Table 5: Details of the simulation study design

Drug	Distribution	Censoring percentage	Median
Control	Weibull (1.3,1.2)	15%	0.9
Experimental	Weibull (1.2,1.8)	27%	1.32
Censoring	Weibull (2,3.1)	-	-

The hazard functions are visualized against time in the next Figure, with the two medians. At the beginning the hazards are close, but later we have divergence.

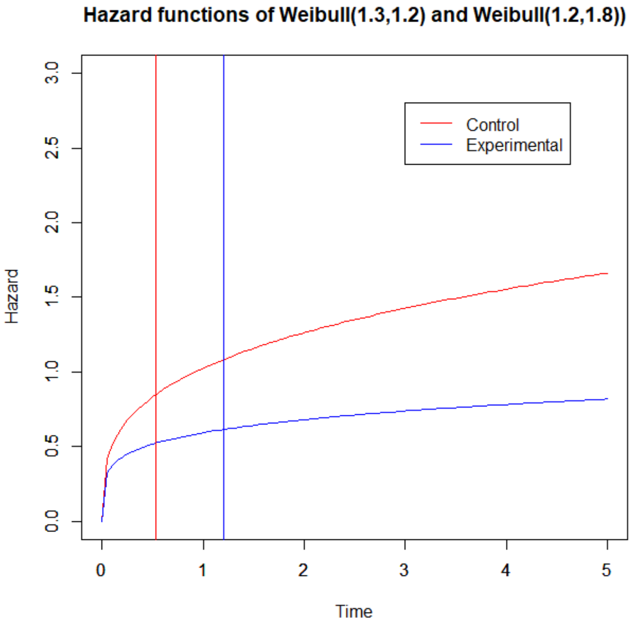


Figure 11: Hazard functions (the curves) and medians (the vertical lines) of Weibull (1.3,1.2) and Weibull (1.2,1.8)

We also visualize the theoretical survival distributions, with the two medians in the next Figure.

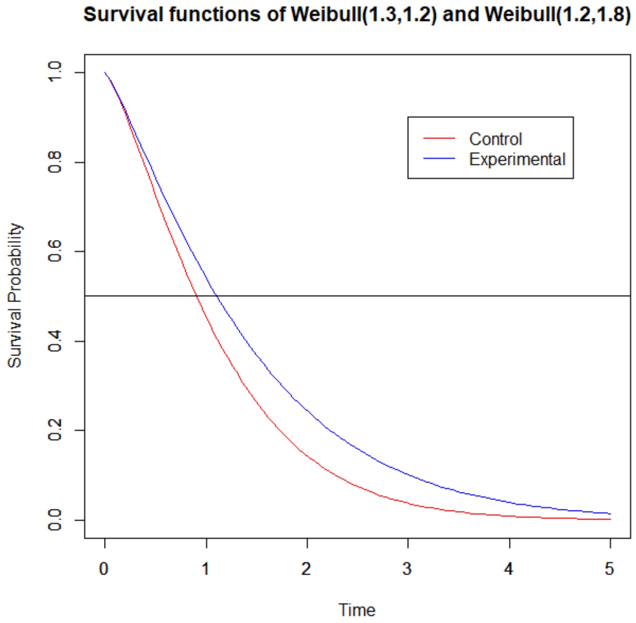


Figure 12: Survival functions (the curves) and medians (the points that the horizontal line crosses each curve) of Weibull (1.3,1.2) and Weibull (1.2,1.8)

The estimated sample size from the procedure with the simulations that we described before is 170. The results with the ssizeCT function are:

Table 6: Results from the powerSurvEpi package

Time	Hazard Ratio	Estimated Sample Size ssizeCT
Control Median (0.9)	0.58	138
Experimental Median (1.32)	0.56	124

In this case the estimated sample sizes are lower than what we actually need. Even if the difference is small when the hazard ratio is 0.58, the test might be underpowered because it uses lower sample than what is needed according to the simulation study.

For the last case (crossing hazards), the simulation design will be:

Table 7: Details of the simulation study design

Drug	Distribution	Censoring percentage	Median
Control	Weibull (0.7,0.9)	16%	0.6
Experimental	Weibull (1.4,1.6)	22%	1.1
Censoring	Weibull (2,3.1)	-	-

For this case, the hazards are crossing in a very important time point. As we can see in the next Figure where the hazards and the two medians are visualized, the crossing point is between the experimental and the control median. Thus, at the beginning of the trial, the control treatment is more efficient. Then the experimental group has a better hazard function. However, as we can see from the theoretical median survival times, the experimental drug is generally more efficient.

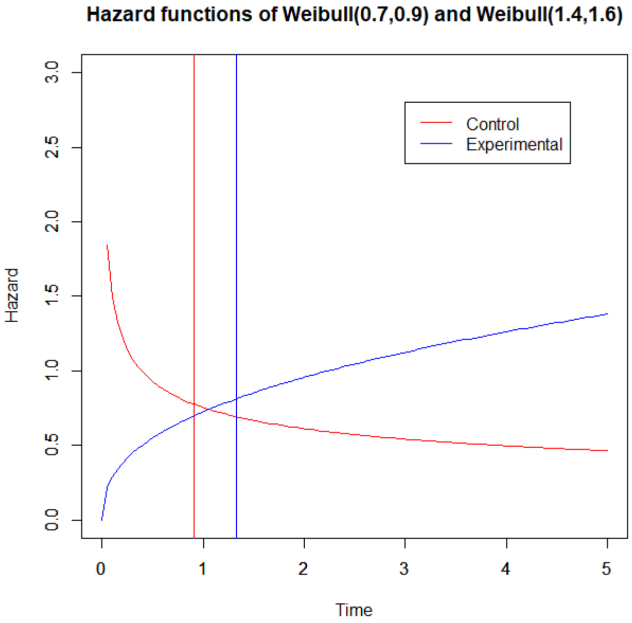


Figure 13: Hazard functions (the curves) and medians (the vertical lines) of Weibull (0.7,0.9) and Weibull (1.4,1.6)

The crossing at the theoretical survival curves, as we see at the next graph, occurs in a time point of no importance. Most of the participants have already experienced the event by then.

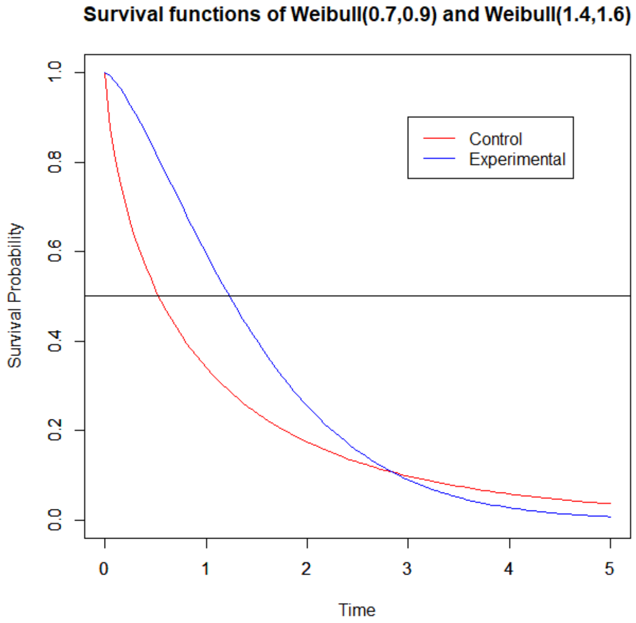


Figure 14: Survival functions (the curves) and medians (the points that the horizontal line crosses each curve) of Weibull (0.7,0.9) and Weibull (1.4,1.6)

The estimated sample size from the simulations is 86. The results from the package ssizeCT function are:

Table 8: Results from the powerSurvEpi package

Time	Hazard Ratio	Estimated Sample Size ssizeCT
Control Median (0.6)	0.61	164
Exper. Median (1.1)	1.11	2986

Both estimations are greater than the actual sample that we need (86). The second one is estimated for a hazard ratio very close to the point of crossing and as a result close to 1. So the sample size is enormously overestimated. Also there is a time point, where the true hazard ratio is 1 and there the null hypothesis is true. It is clear that when the hazards are crossing, the design of the trial can have serious misspecifications.

It is clear that there are many cases where the sample size that we need can be estimated wrongly. It can be overestimated or underestimated and we want to avoid both cases. If we overestimate the sample size that we need, we will enlist more patients and they might be very hard to find or they might increase greatly the cost of the trial. If we underestimate the sample size, then the test will be underpowered. Generally, the miscalculations can be far greater than the ones that we have shown with the simulation studies. In those studies, the only wrong assumption was the assumption about proportionality. If for example, in a clinical trial design, the ssizeCT function is used to estimate the sample size needed, the pilot data set from a previous clinical trial can be very incompatible. The clinicians might think that the previous survival populations match with the current ones, but in fact they do not. So this will be another source of miscalculations. If the clinicians assume exponential distributions in order to estimate the probability of death, then those assumptions can again be very wrong. Thus the final estimation will be also very misleading. All the previous results make clear that the proportionality assumption is very crucial and thus it should be tested in clinical trials. And if it is rejected, the estimated sample size should be questioned as well.

CHAPTER 4: Restricted Mean Survival Time

4.1 Introduction

One common problem that occurs in the context of the survival is that we rarely observe the complete survival times of the participants. As we have already discussed in the survival analysis chapter, in a clinical trial it is almost impossible to observe the death of all the participants, because of the time limitations and the fact that the participants can leave the trial at any time. Thus, estimating the mean survival time is impossible and usually the median survival times are used or another percentile statistic. The restricted mean survival time (RMST), is the mean survival time, until a certain time point. So the interpretation is that a patient's mean survival time, until a certain time point, will be equal the RMST at that time point. This summary measure is very convenient because we can compare two survival curves at any time point we want. Furthermore, there is no need for assumptions about the curves or their relationship in order to be computed. As we will see later, this is a great advantage, because the most popular summary statistic, the HR, needs the assumption of proportional hazards. In this chapter, we will also show a test that tests the significance of the RMST difference of two survival curves. We will also simulate the power of this test, and compare it with the Log Rank test. Also, we will discuss the time point of RMST estimation and how it must be chosen.

4.2 Definition of the RMST

Since the mean survival time is the area under the survival curve (a result that is derived from equation (2.9)), The Restricted Mean Survival Time at time point t^* is the area under the survival curve, up to t^*. So the mathematical expression is:

$$RMST_{t^*} = \int_0^{t^*} S(t)\,dt. \tag{4.1}$$

Where $S(t)$ is the survival distribution. Obviously, for $t^* \to \infty$ we have that $RMST_{t^*} \to \mu$ where μ is the mean survival time and its formula is given by the equation 2.6. Now we give an example of RMST to make it clearer. For the example we take the Exponential(1) distribution, which has mean survival time 1. We will compute the RMST for the t^* values 1, 2 and 10. In Figure 15 we visualize the survival distribution against time and we also visualize the t^* values as vertical lines. For the general case of exponential (λ) those values can be computed by the following formula:

$$RMST_{t^*} = \int_0^{t^*} S(t)\,dt \overset{(2.18)}{\Longleftrightarrow}$$

$$RMST_{t^*} = \int_0^{t^*} \exp(-\lambda t)\,dt = \frac{1}{\lambda}(1 - \exp(-\lambda t^*)) \tag{4.2}$$

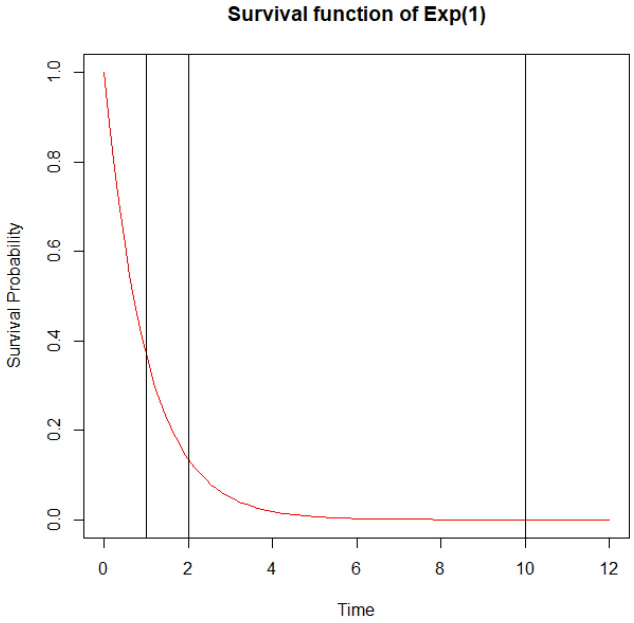


Figure 15: Survival function of Exponential (1) and t^*values 1,2 and 10.

The mean survival is the surface under the survival curve, while the $RMST_{t^*}$ is the surface under the survival curve and until t^*. In the table 9 we give the values of the RMST.

Table 9: RMST values

t^*	$RMST_{t^*}$
1	0.632
2	0.864
10	0.999

Note that as we get further on the axis of time, the closer the RMST gets to the mean survival value, which is a consequence of the asymptotic property that we mentioned before. Also note that the earlier the time of estimation, the lower the value of the RMST.

Now we will visualize in Figure 16 the RMST values of two Weibull distributions. The first will have shape parameter below 1 and the second above 1.

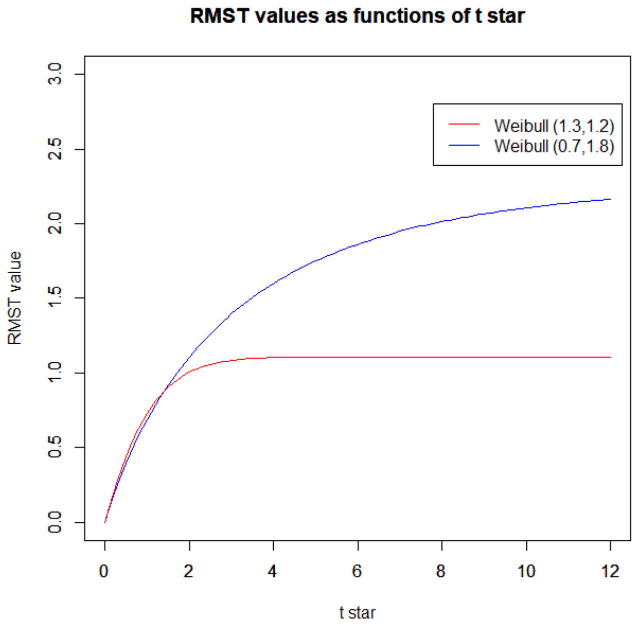


Figure 16: RMST values of Weibull (1.3,1.2) and Weibull (0.7,1.8)

A natural estimator of the RMST is given, if we estimate $S(t)$ with the Kaplan-Meier estimator. The estimator is:

$$\widehat{RMST_{t^*}} = \int_0^{t^*} \widehat{S(t)}\,dt. \tag{4.3}$$

Where $\widehat{S(t)}$ is the Kaplan Meier estimator.

A direct method of computing (4.3) is by integrating the Kaplan Meier estimations of the survival curve, up to t^*, with the following formula from $Wei\ et\ al$ (2015):

$$\sum_{j=1,0<tj\leq tj}^{k} \widehat{S(t_j)}\left(t_{j+1}-t_j\right)+1-(t_1-0). \tag{4.4}$$

Where k is the number of events and t_j the time of the j_{th} event and $\widehat{S(t_j)}$ the Kaplan Meier estimator at that time.

We will now present a different method, using the Jackknife method and pseudo-values, as proposed by *Andersen and Perme* (2010). First of all we will describe the Jackknife method. Suppose we have a sample with size n and θ is the parameter of interest. We then estimate θ from the whole sample and we get $\hat{\theta}$. Then we remove the i_{th} observation and estimate θ from the remaining observations and we have $\widehat{\theta_{-\iota}}$. Then the pseudo-value $\widehat{\theta_\iota}$ is given by the formula:

$$\widehat{\theta_\iota} = n\hat{\theta} - (n-1)\widehat{\theta_{-\iota}}. \tag{4.5}$$

And the pseudo-value estimator is given by the formula:

$$\widehat{\theta_{pseudo}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta_\iota}. \tag{4.6}$$

And the variance of the pseudo-value estimator is given by the formula:

$$var(\widehat{\theta_{pseudo}}) = \frac{\sum_{i=1}^{n}(\widehat{\theta_\iota}-\widehat{\theta_{pseudo}})^2}{n(n-1)}. \tag{4.7}$$

For the specific case of the RMST estimation at time point t^* the pseudo-values are:

$$\widehat{RMST_\iota} = n\int_0^{t^*}\widehat{S(t)}\,dt - (n-1)\int_0^{t^*}\widehat{S_{-\iota}}\,(t)dt. \tag{4.8}$$

Royston and Parmar (2011) proposed a flexible parametric model for modeling the cumulative hazard function that is easy to extend for non-proportional hazard cases. With this estimation we can use the equation 2.15 that transforms the cumulative hazard function to the survival function and then compute the integral of the equation 4.2. The model has the form:

$$ln\,H(t;x) = ln\,H_0(t) + x'b = s(ln\,t) + x'b. \tag{4.9}$$

Where $H_0(t)$ is a baseline cumulative function and x' is a covariate vector. For the modeling of the log cumulative hazard function $ln\,H_0(t) = s(ln\,t)$ we use a restricted cubic spline in log time:

$$s(\ln t) = \gamma_0 + \gamma_1 \ln t + \gamma_2 u_1(\ln t) + \cdots + \gamma_{k+1} \mathrm{u}_k\,(\ln t). \tag{4.10}$$

Where the K+1 basis functions except the first ($\ln t$) depends on an interior knot that connects a pair of cubic polynomial segments of log time. The spline basis functions ($u_1 \dots u_k$) are constructed in such a way that their segments are connected at the knots, and also their first and second derivatives are also connected at the knots. In addition, the spline must be linear in log time in the tails beyond the extremes of the observed event times.

Note that this model provides us a fully parametric estimation of the cumulative hazard function and as a result a fully parametric estimation of the survival function from equation 2.15. As a result the integration is easily done. Also the parametric model gives us the opportunity to extrapolate and compute the RMST with t^* values bigger than the event times that we have observed. The previous estimations that we discussed were using the Kaplan-Meier estimation of the survival function and that is why we cannot estimate the curve beyond the latest event time.

4.3 Test for the significance of the RMST difference and sample size calculation

First of all we will define the RMST difference between two survival curves. It is the difference between the two RMSTs that have been estimated at the same time point t^*. So for two survival curves S_1, S_2 we have:

$$\hat{\Delta} = \int_0^{t^*} \widehat{S_1}(t)\, dt \text{-} \int_0^{t^*} \widehat{S_2}(t)\, dt. \tag{4.11}$$

For a randomized clinical trial the two RMSTs are independent. So the variance of their difference is the summation of the two variances:

$$var(\hat{\Delta}) = var(\widehat{RMST_1}) + var(\widehat{RMST_2}). \tag{4.12}$$

Panagou (2014) in her MCs thesis showed that if we use the Kaplan-Meier estimator for the estimation of the RMST then we have that:

$$\sqrt{n}\left(\widehat{RMST} - RMST\right) \xrightarrow{d} N(0, \sigma^2). \tag{4.13}$$

Where n is the sample size and σ^2 is given by the following expression:

$$\sigma^2 = \int_0^{t^*} \left[\int_t^{\infty} h(u)du\right]^2 \frac{dH(t)}{h(t-)h^C(t-)}. \tag{4.14}$$

Where $H(t)$ is the cumulative hazard function, $h(t-)$ is the hazard function right before the time point t and $h^C(t-)$ is the hazard function without censoring. Since the estimations of the two RMSTs are independent we have for $\hat{\Delta}$:

$$\sqrt{n}\left(\hat{\Delta} - \Delta\right) \xrightarrow{d} N(0, \sigma^2). \tag{4.15}$$

Where $\sigma^2 = \sigma_1^2 + \sigma_2^2$ and σ_1^2, σ_2^2 are the variances of the two RMSTs that are computed with the formula (4.14). If we want to test the hypothesis $H_0: \Delta = 0$ we use the formula (4.15).

Based on the previous test we can calculate the sample size that we need, for a targeted power β and a significance level α. The formula that $Royston\ and\ Parmar$ (2013) gave is:

$$size = \frac{zz^2 * var(\hat{\Delta})}{\Delta^2}. \tag{4.16}$$

Where $zz = z_\beta + z_{1-\frac{a}{2}}$. We can see that the formula does not need any parametric assumptions about the two survival curves or the censoring rate. However it does need an estimation of the variance. So if we use the RMST difference test in a clinical trial, we need some survival data in order to have an estimation of the variance before we calculate the final sample size. That problem can be solved with an adaptation of the clinical trial, as we will see later.

4.4 Choice of the time point (t^*)

The RMST as a summary statistic and the RMST difference test have many advantages. However, the main difficulty is the choice of the time that it will be estimated. Generally, there is not a standard rule for this choice and that can lead to questionable or biased decisions. Now in Figure 18 we will visualize the RMST difference as a function of t^* from two Weibull distributions that have crossing survival curves. In Figure 17 we visualize the curves and also the time that the RMST difference changes sign.

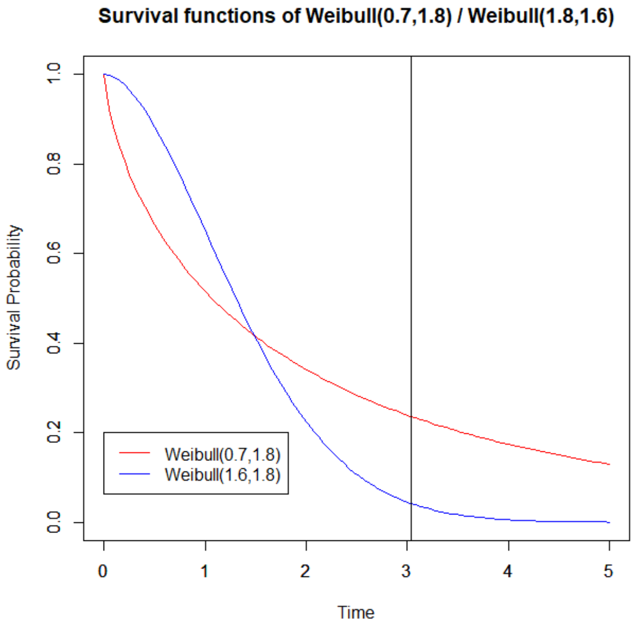


Figure 17: Survival curves of Weibull (0.7,1.8) and Weibull (1.6,1.8). The black line is the time that RMST difference changes sign.

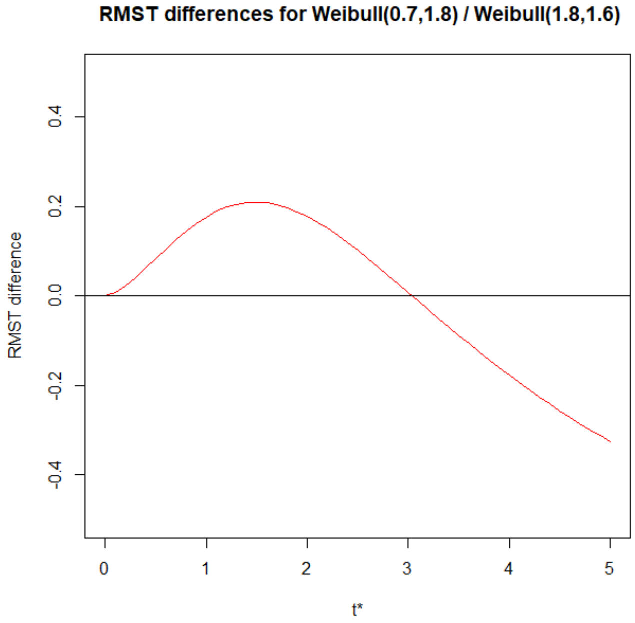


Figure 18: RMST difference as a function of t^*

In Figure 18 we can see that the RMST difference in this particular example changes sign. So for $t^* = 2$ we would have a positive RMST difference and for $t^* = 3.5$ negative RMST difference. That example shows the importance of t^* choice. The rule of this choice, or the specific t^*should be prespecified at the clinical design. As *Hajime et al.* (2014) proposed, t^* should be linked to the clinical relevance and its aims. For example, if we are interested in the early effect of the experimental drug we can choose an early time point for t^*. Also, historical evidence from similar clinical trials must be taken into account, in order to find the optimal t^*.

There is also a different approach for the choice of t^* found in the literature. It proposes that the choice of t^* should minimize the sample size needed, for a given power and significance level. *Royston and Parmar* (2013) presented the ART procedure and proposed an extension of it. The extension does not assume proportional hazards and calculate the sample size based on the RMST difference. In the ART procedure, we set an accrual period of K_1 units of time and a follow up period of K_2 units of time. Then we have K_1+K_2 = K as a study time. We then fit a parametric model, with the patients recruited at K_1 and estimate the variances of the two RMSTs. Then we simulate survival times, for each group based on the previous estimations and calculate the sample size needed for given power and significance level. This procedure is repeated for different t^* values in the range of $(K_1, K_1{+}K_2)$ and we choose the one that minimizes the sample size. One major disadvantage of this method, is that it needs a parametric model before we get any outcome from the clinical trial. We need to model both the accrual rate and the survival curves and then estimate their parameters. So if the choice of models is not correct, the variance estimate will be false and all the other calculations as well.

Another procedure also proposed by *Royston and Parmar* (2013) is the data maturity analysis. This procedure does not assume any parametric model at the beginning of the trial, so there is not the problem of goodness of fit. There is a formula that calculates the “percentage of maturity of the data” and this is

$$pmat = \frac{100*\Delta^2}{zz^2*var(\hat{\Delta})}. \tag{4.17}$$

Where $zz = z_\beta + z_{1-\frac{a}{2}}$ and β = the targeted power and a = the significance level. Δ is the RMST difference that we want to detect. So this formula is a function of $\hat{\Delta}$ which is function of the sample size and t^*. When pmat $= 1$, then the targeted power has been achieved and the RMST difference is ready to be estimated. Therefore, by varying t^* over plausible values, or the sample size, we can stop the clinical trial by that criterion. Or else, we can calculate pmat periodically, until it is equal to 1 and then stop the trial.

Another method for choosing t^*, is to set it equal with the minimum of the maximum observed event, of the two groups. So for one out of two groups we will estimate the RMST using all the available information. On the other group, we will lose information, but it would be the minimum loss that we can achieve without extrapolating the estimation of the other group. This strategy is a safe option to use as much information as possible, without the risks of extrapolation and seems like a sensible, data driven choice.

4.5 Comparison between RMST difference test and Log Rank test

In this section, we will compare the RMST difference tests and the Log Rank test, in terms of sample size needed and power. The main issue with the “popular” Log Rank test is its performance under non proportional hazards. In section 3.5, we study the difference of the estimated sample size needed, when the estimation was done under the assumption of proportional hazards (with the $powerSurvEpi$ package) and when it was done by simulation. The estimation through simulations was done in the following way. For different sample sizes, we estimated the power for given significance. By this procedure, we could estimate the sample size needed to achieve power 80% on 10% significance level and with a study time equal to 1.3 times the median of the experimental drug. That simulation study, took part under different types of violation of the proportionality assumption. So we already have the results, from 3 types of violation that connects a certain sample size with power equal 80%, on a significance level 10%. From the same hypothetical clinical trials (same survival and censoring distributions) and with the sample size that achieves 80% power on 10% significance level with the Log Rank test, we will estimate the power of the RMST difference test. Of course we will also add a case where the proportionality assumption does hold, because it is crucial to compare the two tests in situation where the Log Rank test is used correctly. Additional to the Log Rank test, in the cases of non-proportional hazards we will also examine the power that we have from the Fleming-Harrington G(ρ) tests (The family was presented at section 2.6). The ρ will be 0.2 and 0.8 in order to investigate a wide range of values. The cases that we will examine will be:

1) Proportional hazards
2) Early diverging hazards
3) Late diverging hazards
4) Crossing hazards

Since the RMST difference test has no parametric assumptions, or assumptions about the relationship of the hazards, we are not afraid that a violation can happen. However, as we discussed before, it is very important to choose a suitable t^*. We are interested to estimate the power of the RMST difference test for different t^*. Thus, for each case, we will choose an “early” and a “late” t^* for estimating the power. The early t^* will be the 0.8 times the experimental median and the late will be 1.3 times the experimental median, which is also the total study time. By estimating those 2 powers, we will compare not only the tests, but also the performance of the RMST difference test for different designs. For the estimations of the RMSTs and their variances we will use the R function rmsth from the package $PWEALL$.

For the first case (proportional hazards), the simulation design will be:

Table 10: Design of the simulation

Drug	Distribution	Censoring percentage	Median
Control	Exponential (1)	14%	0.69
Experimental	Exponential (1.5)	24%	1.03
Censoring	Weibull (2,3)	-	-

We will visualize the hazard functions against time and we will add the two medians (Figure 19). The two hazard functions are proportional.

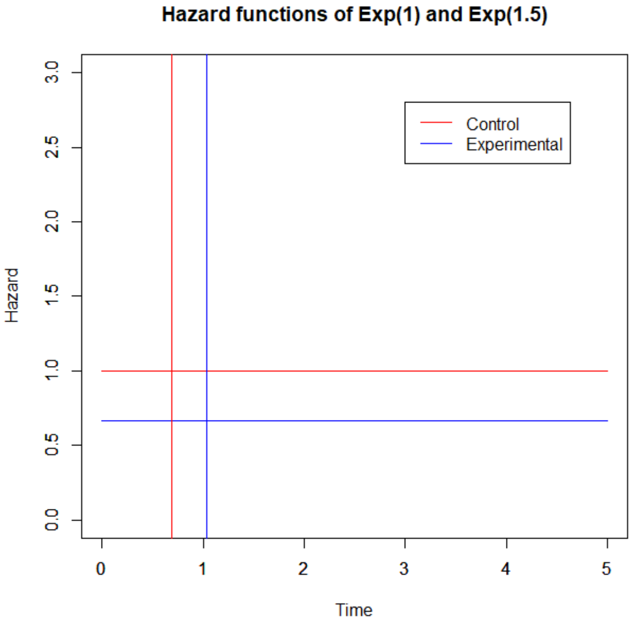


Figure 19: Hazard functions (the curves) and medians (the vertical lines) of Exponential (1) and Exponential (1.5)

We also need to visualize the two theoretical survival curves in Figure 20 in order to understand how two survival distributions with proportional hazards look like. The 2 medians have been also added by drawing a vertical line crossing the y-axis at 0.5.

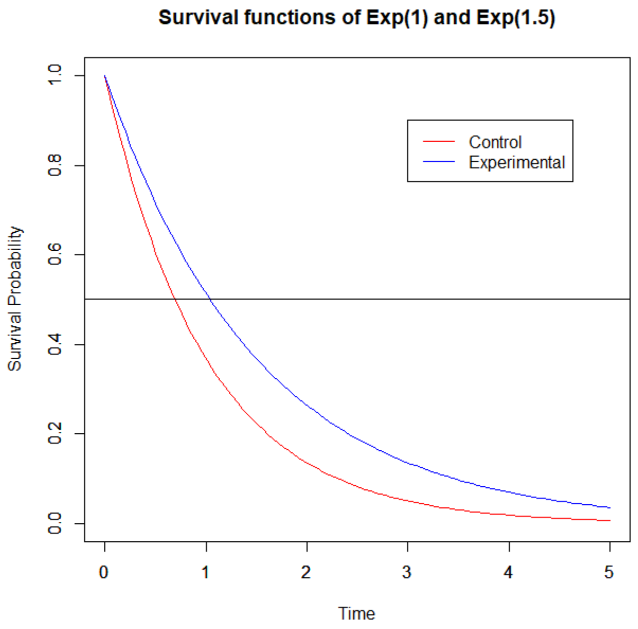


Figure 20: Survival functions (the curves) and medians (the points that the horizontal line crosses each curve) of Exponential (1) and Exponential (1.5)

The simulated sample size needed for power 80% and significance level 10%, when we use the Log Rank test, is 238 (equal allocation to groups). The simulated powers are:

Table 11: Results of the simulation study

TEST	SIMULATED POWER
Log Rank	80%
RMST dif. Early $t^* = 0.831$ (0.8*Exp. Median)	75.8%
RMST dif. Late $t^* = 1.351$ (End of trial)	84.4%

We can see that the choice of early t^* leads to loss of power compared to the Log Rank test. However, when we choose the end of the trial as t^*, we gain power. So in this example, the RMST difference test can increase the power, if we choose a late t^*.

For the second case (early diverging hazards), the simulation design will be the one described in table 3 in section 3.5. In Figures 9 and 10 we can see the visualization of the hazard functions and the survival functions respectively. If we use the Log Rank test, on a significance level 10%, the estimated sample size that we need to achieve power 80% is 114. The results are:

Table 12: Results of the simulation study

TEST	SIMULATED POWER
Log Rank	80%
RMST dif. Early $t^* = 0.884$	93.4%
RMST dif. Late $t^* = 1.436$	91.5%
G(ρ) $\rho = 0.2$	82.5%
G(ρ) $\rho = 0.8$	85.9%

It is clear that with the same sample size, the RMST difference test is more powerful than the Log Rank test. Because of the early divergence of the hazards, even at early time points the proportionality assumption is heavily violated. Thus, at any time point the RMST difference test is more appropriate. Now for the G(ρ) tests it seems that they do increase the power. That happens because in early diverging hazards, there is great difference at the beginning in the two populations in terms of event numbers. So the G(ρ) family that gives more weight in the early events take advantage of this and detect the difference of the two curves more easily. That is why the power is greater for $\rho = 0.8$ than for $\rho = 0.2$ because in the first case the weight to the early observations is greater. However the increase that we get from the G(ρ) family is not so great as that from the RMST difference test.

For the third case (late diverging hazards), the simulation design will be the same as in table 5 in section 3.5. At Figures 6 and 7 we can see the hazard functions and the theoretical survival distributions. The estimated sample size that achieves 80% power on 10% significance level, when we use the Log Rank test is 170. The fixed t^* values will be 1 and 2. The results are:

Table 13: Results of the simulation study

TEST	SIMULATED POWER
Log Rank	80%
RMST dif. Early $t^{*} = 1.061$	68.1%
RMST dif. Late $t^{*} = 1.724$	83.4%
G(ρ) $\rho = 0.2$	78.7%
G(ρ) $\rho = 0.8$	80.3%

In this case we can see that the power of the Log Rank is better than that of the RMST difference test for the early t^{*}. However the RMST difference is slightly better for late t^{*}. That is happening because of the late divergence between the hazards. At early time points the hazard functions are almost equal. However, as the time points increase, the difference between the hazard functions is more detectable. That is why the RMST difference test performs best at the furthest time point, which is the end of the trial (late t^{*}). The G(ρ) achieves almost equal power for both ρ values. Those results are due to the fact that the violation of the proportionality is not so heavy because the two shape parameters of the simulated populations are very close. However, the RMST difference test with late t^{*} seems again to be beneficial.

For the fourth case (crossing hazards), the simulation design will be the same as in table 7 in section 3.5. In Figures 13 and 14 we can see the hazard functions and the theoretical survival curves respectively. The sample size needed that was estimated previously is 86. The results are:

Table 14: Results of the estimations study

TEST	SIMULATED POWER
Log Rank	80%
RMST dif. Early $t^{*} = 0.985$	98.9%
RMST dif. Late $t^{*} = 1.6$	97.2%
G(ρ) $\rho = 0.2$	84.3%
G(ρ) $\rho = 0.8$	92.1%

It is clear that the RMST difference test is more powerful than the Log Rank test, at every t^*. Its power is almost 1, which means that we could get a sufficient power with a smaller sample size. So the RMST difference test with the two fixed t^* values, capture this superiority quite easily. G(ρ) family also increases the family, especially in when $\rho = 0.8$, but it is not so efficient as the RMST difference test. In that example we had the most important violation of the proportionality assumption which is when the hazards are crossing. In this case we can see that all the alternatives to the Log Rank test that were tested are better than it.

Through the simulation study, we examined different cases of violations of the proportionality assumption and also a case where the assumption holds. The RMST difference test, with t^* set to be the end of the trial was always more powerful than the Log Rank test, even in the case of proportional hazards. However, the early t^* values were sometimes less powerful (proportional hazards and late diverging hazards). Thus, the safest choice seems to be to set as t^* the end of the trial. Also the early t^* values do not use the whole available information. That is why there must exist a clinical related purpose for an early t^* value to be chosen (e.g. a disease that mostly affects older people and we are interested in an early superiority of the experimental drug, while we are not so interested in the later events). *Huang and Kuan* (2018) compared the RMST difference test and the Log Rank test under many proportional and non-proportional cases and found that when choosing a late t^*, in most of the cases the RMST difference test is more or equally powerful to the Log Rank. The G(ρ) family was more or equally powerful with the Log Rank test in all of the cases of non-proportional hazards. However the RMST difference test with late t^* was always more powerful than both ρ values that were used.

Since the proportionality assumption is very restrictive and almost never holds, the RMST difference test with t^* set to be the end of the trial is a reliable alternative to the Log Rank test. The G(ρ) family, even if it seems to increase (or at least not decrease) the power in cases of non-proportionality, has an interpretation problem.

While the Log Rank tests the hazard ratio, when we add the weight function of the G(ρ) family it is not quite clear what we really test. Because we give more weight to some observations and the statistic that we produce is not so clearly related to the hazard ratio. Instead, the RMST difference test has a very natural interpretation. For all the above reasons, the RMST difference test with t^* set to be the end of the trial will be the test that we will use for the adaptive design, as an alternative to the Log Rank test.

CHAPTER 5: Adaptive Design

5.1 Aim of the adaptive design

The problem that we want to solve with the adaptive design is the problem of the proportionality assumption that is often violated. The main problem of this assumption is that it is very strict and as a result it is often violated. However, it is used very widely because of the simplification that provides. The Log Rank test can be used and the hazard ratio can be reported as summary statistic. However, as we saw at the second chapter (Clinical Trials), the violation of the proportionality can lead to serious underestimation and overestimation of the sample size needed. Therefore we might have power problem or we might enlist more patients than the number we need. Also, there is problems with the final summary statistic that we use. Since we report the hazard ratio, if the proportionality assumption does not hold, the hazard ratio is a function of time, and thus it is meaningless to report a single value for it.

The adaptation that we propose tries to solve this problem. Since we cannot know if the proportionality assumption holds at the beginning of the trial, we have to test it at some point, when we will have already gathered some survival data. In the third chapter (Restricted Mean Survival Time) we saw that the RMST difference test performs better or equally with the Log Rank test, when the proportionality assumption is violated. So the main idea is that we want to begin the design with the standard set up, because it provides us with many formulas for power and sample size calculation. At a predefined point we will test the proportionality assumption and if we do not reject it we will continue with the standard set up. If the assumption is rejected, then we will re-design the trial based on the RMST difference test. Since we have that it performs equally or better than the Log Rank test, we will estimate again the sample size that we need based on that. The adaptation will fix the power and sample size problems that might occur and at the end of trial we will report the RMST difference instead of the hazard ratio.

5.2 Example of adaptive design

In this section we will give an example of a clinical trial presented by Chow and Chang (2008). The reason that we chose this design is because it illustrates an idea that we are also going to use. It is the possible re-estimation of the sample size based on interim analysis with gathered data.

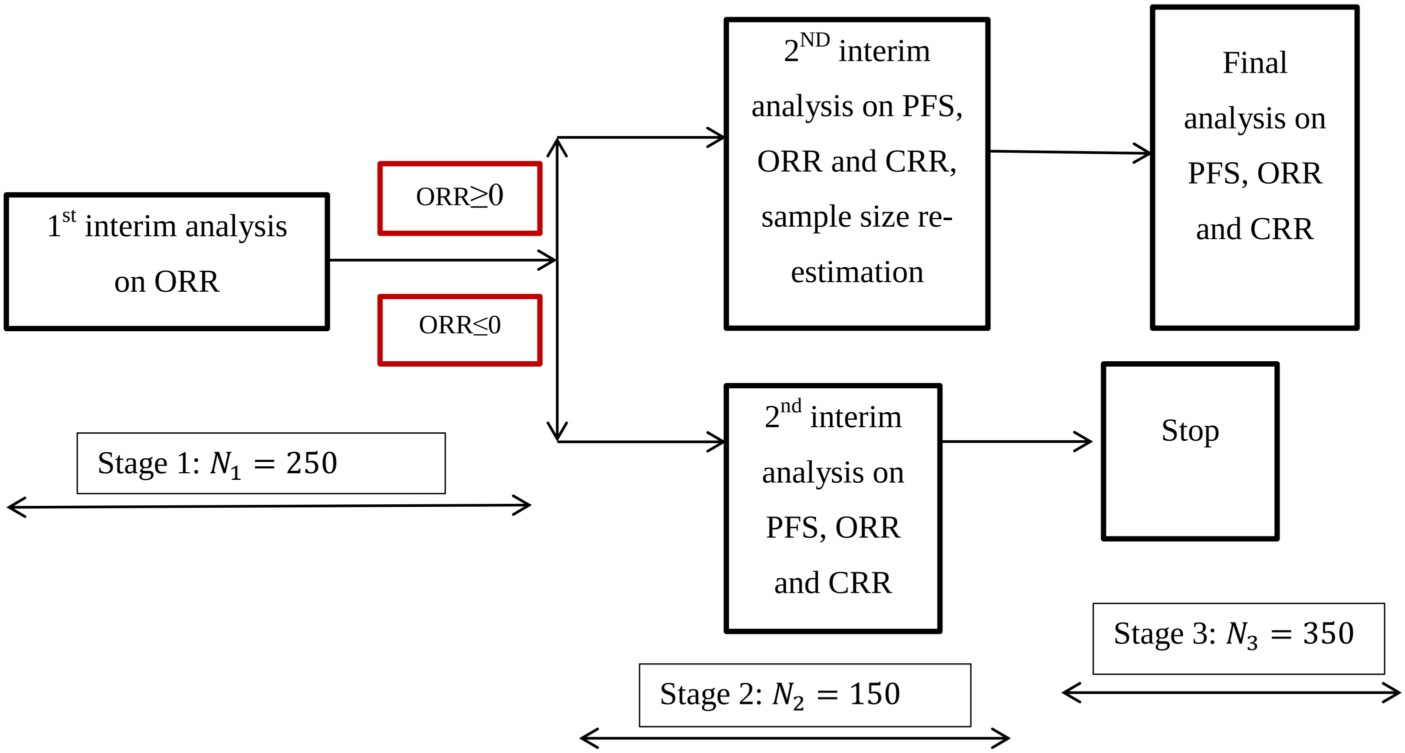
The trial is a phase III with two groups (test-control) and it tests cures for Non-Hodgkin's Lymphoma. The primary endpoint is the progression-free survival time (PFS). There are also secondary endpoints which are the overall response rate (ORR) and the complete response rate (CRR).

We assume a uniform accrual period of 9 months and a follow up period of 23 months. The sample size is 375 subjects per group. However, we split it into three stages. In the first stage, each group has 125 subjects and we conduct interim analysis with them based on ORR. If the difference (test-control) is $\Delta_{ORR} > 0$ then we continue enlisting patients. Else, the accrual is stopped prematurely, and one final analysis for efficacy based on the PFS is conducted. Also we examine possible efficacy based on the secondary endpoints.

If the enlisting continues, then we conduct a second interim analysis based on PFS. From this analysis we will claim either efficacy, or futility or we will go to the next step and we might do sample size re-estimation. In the last stage, we will do the final analysis on PFS and if we find significant results, we will continue testing the secondary endpoints for possible efficacy.

It is clear that based on the result of the interim analysis on ORR, the final sample size can vary from 250 to 750 patients. Now we will visualize the process in Figure 21.

Figure 21: Non-Hodgkin's lymphoma trial designed by *Chow and Chang* (2008)


1st interim analysis on ORR
ORR≥0
ORR≤0
2ND interim analysis on PFS, ORR and CRR, sample size re-estimation
Final analysis on PFS, ORR and CRR
2nd interim analysis on PFS, ORR and CRR
Stop
Stage 1: $N_1 = 250$
Stage 2: $N_2 = 150$
Stage 3: $N_3 = 350$

5.3 Description of the adaptive design

The adaptive design is initially based on the Log Rank test. Therefore, as in a typical design with the proportionality assumption, we have to determine the targeted hazard ratio that we want to detect and also the power and the significance level. Then, we will assume parametric forms for the two survival curves and a censoring rate, based on historical data and based on that we will calculate the sample size needed as described in Section 3.3. However, if the adaptation is needed, the test that we are going to use is the RMST difference test. So we also need to determine the targeted RMST difference that we want to detect with the same power and significance level. We would like to be able to relate the hazard ratio with a RMST difference at each t^* value.

Unfortunately, that cannot happen because pairs of distributions with the same hazard ratio produce different RMST differences and the targeted RMST difference will not be related to the targeted hazard ratio. To illustrate this we give the RMST difference produced by 3 pairs of distributions that have hazard ratio ½ in Figure 22. The cases are: 1) Exp(0.5)-Exp(1) (red curve), 2) Exp(1)-Exp(2) (blue curve), 3) Exp(2)-Exp(4) (purple curve).

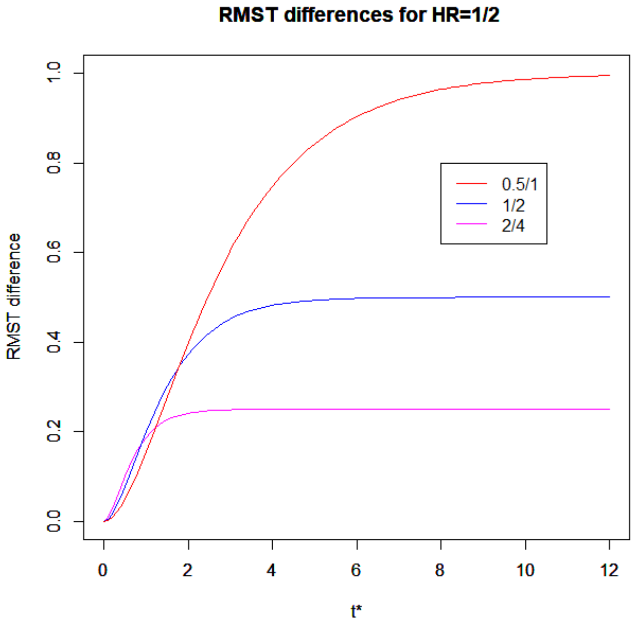


Figure 22: RMST differences as function of t^* produced by pairs of survival distributions with HR=1/2

Since we might use the RMST difference test we must also define t^*, or a decision rule for it. At the third chapter (Restricted Mean Survival Time) we discused the issues about the choice of t^* and whether or not it must be fixed or chosen with a decision rule. At the adaptive design it can also be related to the time point of the testing for the proportionality as we will see later.

The main idea of the adaptive design is that we have to test the proportionality assumption. The test that we are going to use was proposed by $Grambsch\ and\ Therneau$ (1994) and was presented in section 2.7. So we have to find a time point to check it. As we have described in the second chapter (Clinical Trials), the patients of the trial are not enlisted all together, but during a predefined accrual period. After the accrual period is over, then there is a maximum follow up period. We propose that after a predefined portion of the initial sample size has been achieved the assumption must be tested. If the assumption is not rejected the process should continue with the initial design. So in the design we should split the initial sample size in what we will call the “first sample” which will be used for the test and the remaining “second sample”.

If the assumption is rejected then the design must adapt to the RMST difference test. The new sample size should be estimated again with the formula 4.16. If the new sample size needed is estimated lower than the “first sample” then we already have the targeted power and we do not need to enroll more patients. Therefore we can test the RMST difference immediately. If the new sample size needed is estimated bigger than the “first sample” but lower than the initial sample size then we will enlist the remaining patients and continue with the test. However there is also the possibility that the new sample size will be bigger than the initial sample size and there is also possible that the difference will be huge. So at the beginning we have to set a maximum sample size and if the sample size ends up bigger than that we will just use the maximum size. Unfortunately at this case we will have power lower than the targeted. Note that the new sample size estimation is a function of t^* .

At the end of the trial we must report a summary statistic with a confidence interval and also the p-value of the test that we did. If the trial does not adapt to the RMST difference test, we will report a value for the hazard ratio and a confidence interval and the p-value of the Log Rank test. Else we will report the RMST difference with a confidence interval and the p-value of the RMST difference test.

Now we will summarize the design:

Step 0:

1) Assume parametric forms of the two survival curves that lead to proportional hazards

2) Assume a censoring rate for both arms, or one for each arm.

3) Determine the targeted HR, the targeted power and the significance level
4) Estimate the initial sample size needed for the Log Rank test
5) Split the initial sample size into “first-second sample”
6) Determine the accrual and the follow up period
7) Determine the targeted RMST difference
8) Determine t^* or a decision rule for it
9) Determine the maximum sample

Step 1:

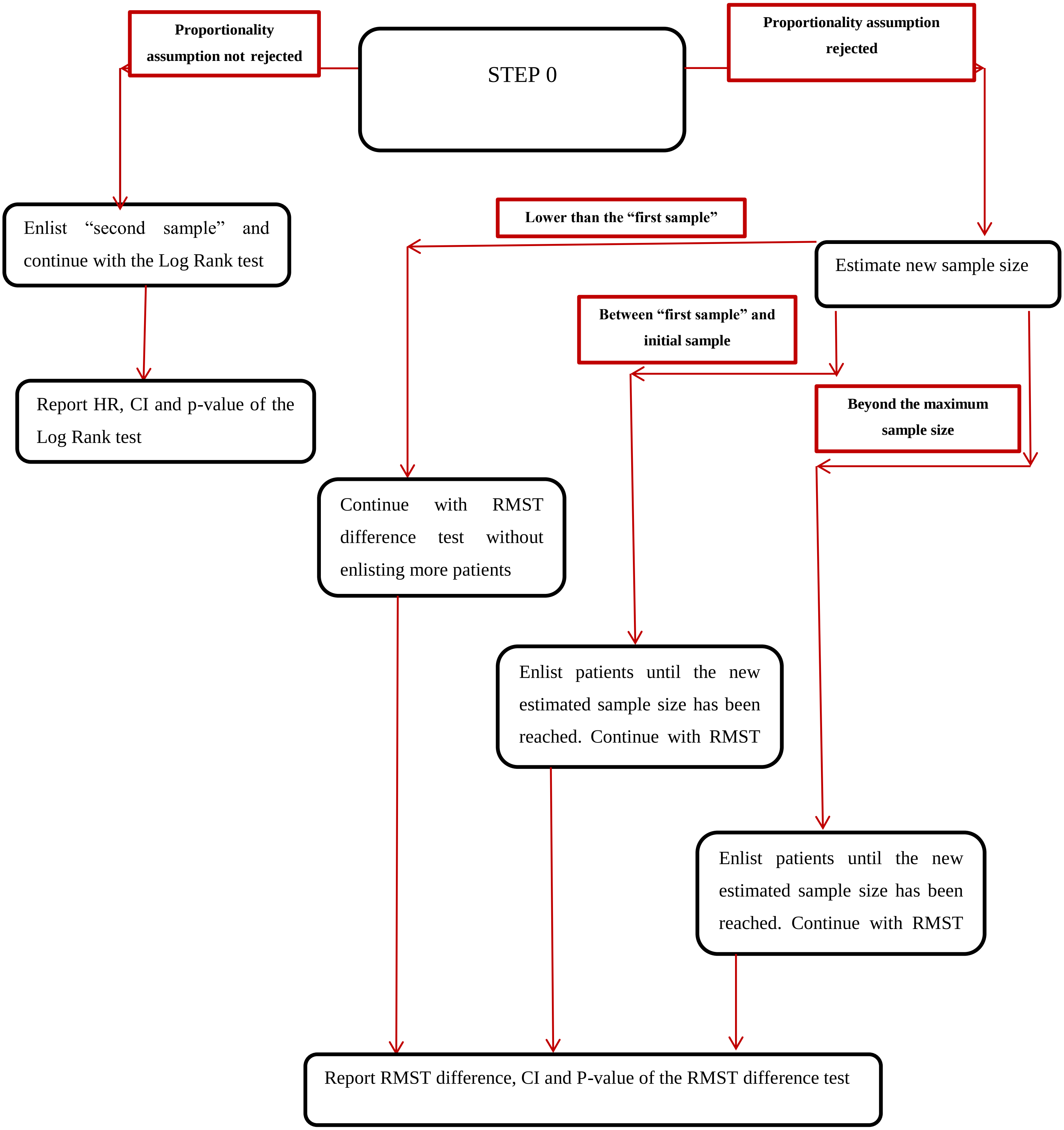
1) Test the proportionality assumption on the “first sample” with the proportionality test described in section 2.7.
2) If it is not rejected then enlist the “second sample”
3) Continue to the follow up period
4) Report the HR with a confidence interval and the p-value of the Log Rank test
5) If it is rejected go to step 2

Step 2:

1) Estimate the new sample size with the formula 4.16.
2) If it is lower than the “first sample” then test immediately the RMST difference
3) If it is between the “first sample” and the initial sample then enlist the remaining patients, continue with the follow up period and test at the end the RMST difference
4) If it is above the maximum sample size, enlist patients until the maximum sample size and continue with the follow up period and the RMST difference test
5) Report the RMST difference and a confidence interval and the p-value of the RMST difference test

We will also add a flow chart (Figure 23) to make it clearer. The first step is whole at the first box because all the procedures in it must be done before the trial begins. After that, we split the next steps is an algorithmic way:

Figure 23: The adaptive design as a flow chart


STEP 0
Proportionality assumption not rejected
Proportionality assumption rejected
Enlist "second sample" and continue with the Log Rank test
Report HR, CI and p-value of the Log Rank test
Estimate new sample size
Lower than the "first sample"
Between "first sample" and initial sample
Beyond the maximum sample size
Continue with RMST difference test without enlisting more patients
Enlist patients until the new estimated sample size has been reached. Continue with RMST
Enlist patients until the new estimated sample size has been reached. Continue with RMST
Report RMST difference, CI and P-value of the RMST difference test

5.4 Simulation study

Now we will test the performance of the adaptive design and compare it to that of the typical design (log rank test and assumption of proportional hazards) in different cases of proportional, non-proportional and crossing hazards. What we are interested for is the power and the mean sample size that the adaptive design produces. For estimating the power we will simulate 1000 survival data sets, from clinical trials that compare two drugs with equal allocations to each group and count the percentage that the difference between the two drugs was correctly detected. The general set up is that we have 2 simulated survival groups that get two different drugs and are compared to each other, with a censoring distribution that censors about 20% of the observations. Firstly, we want to find the sample size that we would need to achieve 80% power at 10% significance level with the typical design for the same cases. For this estimation, for different sample sizes we will generate 1000 clinical trials with the typical design and we will estimate the power, until we end up with a sample size that gives us 80%. We will call this sample size $N_{\log rank}$ as it is the sample size that the log rank test needs to achieve 80% power at 10% significance level. That enables the skipping of assuming parametric forms for the two curves and the censoring distribution and the setting of a targeted hazard ratio to be detected. Now we set the $n_1 = 0.5N_{\log rank}$, $n_2 = 0.5N_{\log rank}$ and $n_m = 1.1N_{\log rank}$. As we saw in section 4.5, the best alternative to the Log Rank of the ones that we studied is the RMST difference test with a late t^*, so our choice for this will be the end of the trial. The time points of the design will be a function of the largest median of the two survival populations. More specifically, the time of testing for the proportionality assumption will be $0.8 * largest\,median$ and the end of the trial (which will also be t^*) will be $1.3 * largest\,median$. The test for the proportionality will be set at 10% significance level. The targeted RMST difference that will be used for the sample size re-estimation will be the real difference between the two RMSTs. At the end of the simulation, we will take the mean sample size that the adaptive design used and we will denote it $N_{adaptive}$. The mean sample size is the mean sample produced from the simulated trials. Note that with the adaptive design, we do not know from the beginning the final sample size that will be used.

We are interested in the relationship between $N_{\log rank}$ and $N_{adaptive}$. Also we will estimate the power of the adaptive design and compare it to the power of the typical design. Now in the next table we give the cases that we are going to test.

Table 15: Design of each simulation study

	CASE	Control/Experimental	Censoring distributions	Time of proportionality testing	End of the trial-t^*
Prop. Hazards	1	Weibull(1,1.1)/Weibull(1,1.8)	Weibull(2,3.3)	0.998	1.621
	2	Weibull(0.9,0.9)/Weibull(0.9,1.8)	Weibull(2,3.2)	0.958	1.557
	3	Weibull(1.1,1.1)/Weibull(1.1,1.8)	Weibull(2,3.2)	1.031	1.676
Non-Prop. Hazards	4	Weibull(1.3,1.2)/Weibull(1.2,1.8)	Weibull(2,3.1)	1.061	1.724
	5	Weibull(0.9,0.9)/Weibull(1.2,1.5)	Weibull(2,2.9)	0.884	1.436
	6	Weibull(1,1.1)/Weibull(1.2,1.8)	Weibull(2,3.2)	1.061	1.724
Crossing Hazards	7	Weibull(0.7,0.9)/Weibull(1.4,1.6)	Weibull(2,3.1)	0.985	1.6
	8	Weibull(0.8,1.3)/Weibull(1.2,1.8)	Weibull(2,3.6)	1.061	1.724
	9	Weibull(0.9,1.1)/Weibull(1.3,1.5)	Weibull(2,3)	0.905	1.47

Now in Figures 24, 25 and 26 we give the hazard functions of the pairs that will be tested. With red curve is the hazard of the control group and with blue the experimental.

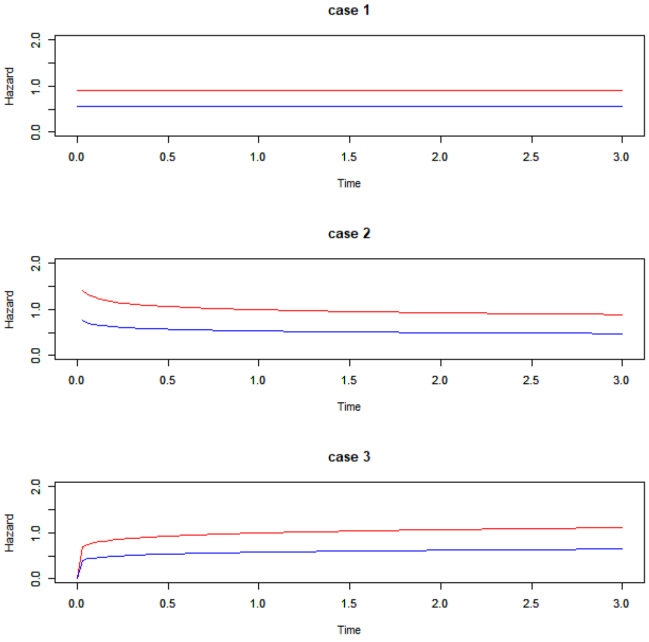


Figure 24: Hazard functions of the proportional cases

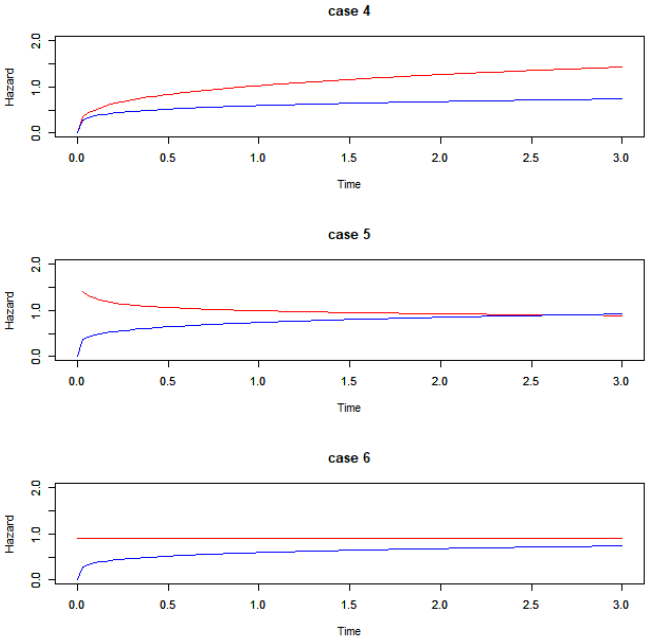


Figure 25: Hazard functions of the non-proportional cases

Note that in case 5 the hazards are crossing. However the crossing is of no importance, because it takes place at a moment when most of the patients have already died.

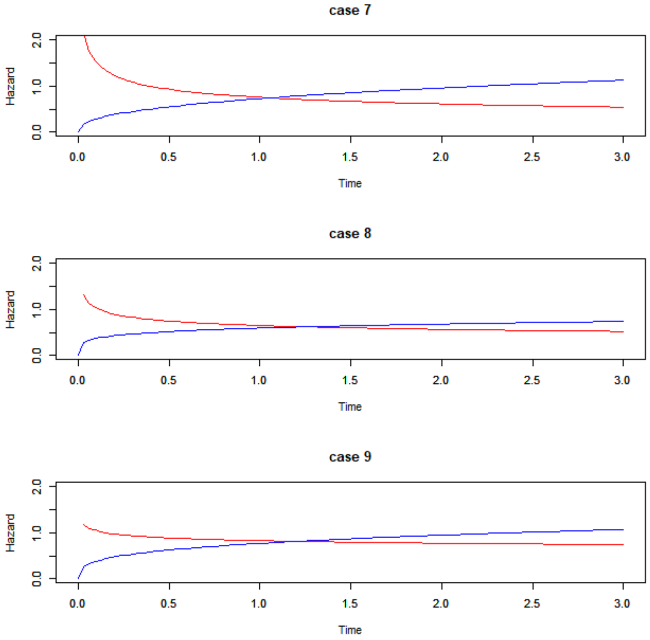


Figure 26: Hazard functions of the crossing hazards cases

Now in Figure 27 we give the survival functions of each pair that we are going to test. Again, the red curve is for the control group and the blue curve for the experimental.

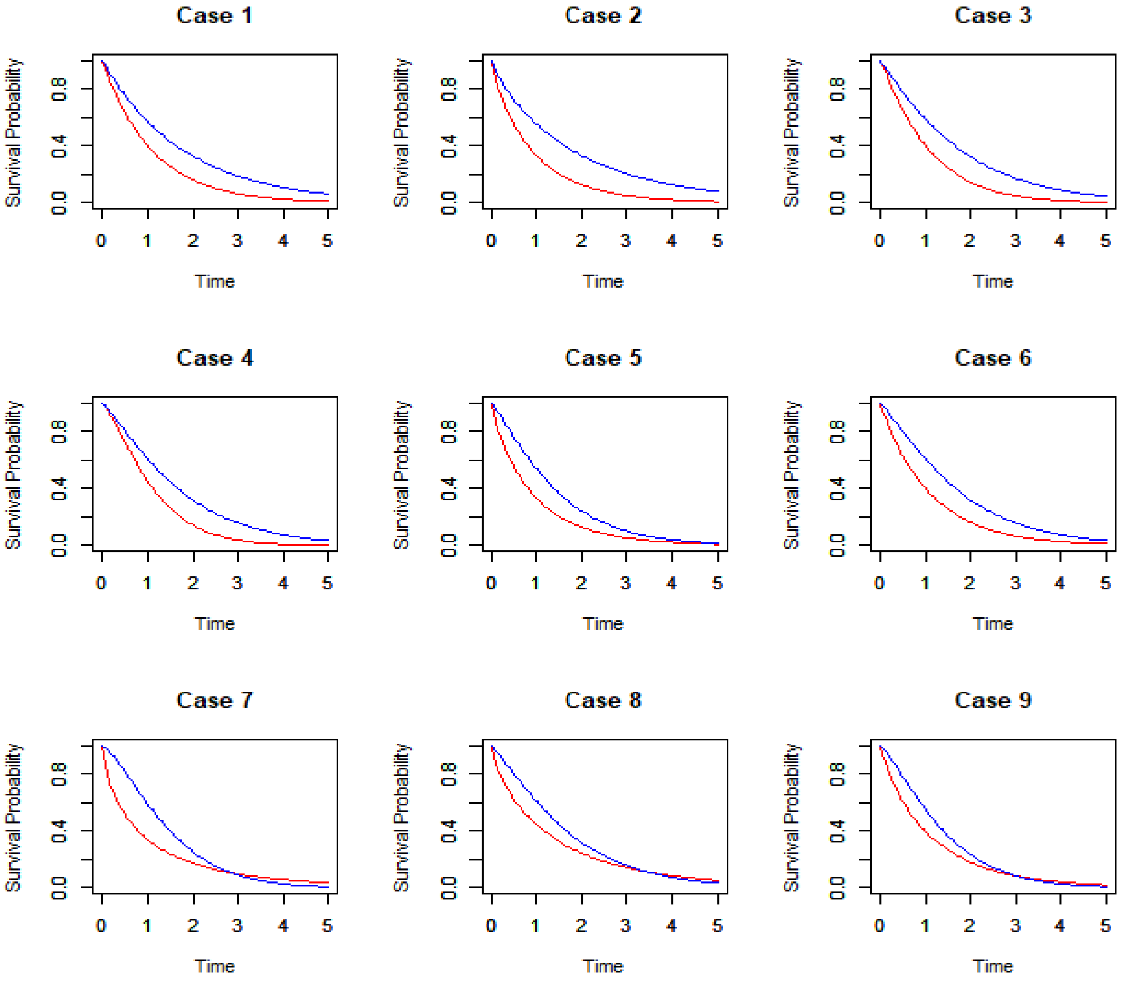


Figure 27: Survival curves of each pair of the simulation study

The results of the study are shown in table 16:

Table 16: Results of the simulation study

	Case	Control/Experimental	% OF RMST TEST USE	$N_{\log rank}$	$N_{adaptive}$	POWER	$\frac{N_{adaptive} - N_{\log rank}}{N_{\log rank}}\%$
Prop. Hazards	1	Weibull(1,1.1)/Weibull(1,1.8)	10.1%	160	151.76	80.5%	- 5.15%
	2	Weibull(0.9,0.9)/Weibull(0.9,1.8)	11.1%	100	85.7	81.5%	- 14.13%
	3	Weibull(1.1,1.1)/Weibull(1.1,1.8)	10.2%	130	122.837	81.2%	- 5.51%
Non-Prop. Hazards	4	Weibull(1.3,1.2)/Weibull(1.2,1.8)	17.5%	170	158.831	81.3%	- 6.57%
	5	Weibull(0.9,0.9)/Weibull(1.2,1.5)	16.3%	114	104.025	82%	- 8.75%
	6	Weibull(1,1.1)/Weibull(1.2,1.8)	13.3%	116	108.286	80.3%	- 6.65%
Crossing Hazards	7	Weibull(0.7,0.9)/Weibull(1.4,1.6)	39.4%	86	69.058	82.3%	- 19.7%
	8	Weibull(0.8,1.3)/Weibull(1.2,1.8)	39.9%	230	184.115	84%	- 19.95%
	9	Weibull(0.9,1.1)/Weibull(1.3,1.5)	35,1%	228	187.986	83,3%	- 17,55%

From the simulations it is obvious that the adaptive design reduces on average the sample size that we would use with the typical design. Even in cases with proportional hazards, where the adaptation take place only at 10% (type I error on the testing for the proportionality), the sample size has a reduction. In cases of non-proportional hazards that are not crossing (at least during the trial's period), the adaptation takes place in under 20% of the repetitions. That happens because the test for the proportionality is not so powerful in cases where the proportionality is not heavily violated. However, there is a reduction to the sample size as well. In the cases of crossing hazards where the proportionality assumption is heavily violated, the sample size reduction is greater and it reaches almost 20%. All those reductions to the sample size do not reduce the power of the design. In fact, in the cases of crossing hazards we even have a small increase in power. In this simulation study, we were beginning by knowing the exact sample size for achieving 80% with the log rank test. But in real life trials, this size is estimated based on assumed parametric forms for the survival curves and the censoring distribution and of course the assumption of the proportionality of hazards. All these assumptions can lead to under or over powered designs, which means to estimate a sample size that is smaller or bigger than it actually needs for the targeted power. The adaptive design however seems to calibrate this sample size in order to achieve the targeted power. In the simulations we had reduction of the size without losing power. So in cases of underestimation, we are confident that the design will increase the sample in order to achieve the power and in cases of overestimation the design will reduce the sample size. Another advantage of the adaptive design is that when the adaptation is needed then it also changes the summary statistics that we are reporting at the end of the trial. If the proportionality assumption does not hold then not only the use of the Log Rank test is wrong but reporting a single hazard ratio is also wrong since it is not constant. The adaptive design prevents that from happening, because if it detects non-proportionality, then it switches the final summary statistic to RMST difference, which is irrelevant to the hazard ratio.

Chapter 6: Discussion and Further Research

In the present thesis we developed an adaptive design which tests the proportionality assumption at the first stage and then deciding on this selects an appropriate test statistic to examine the difference between the two arms. We have shown through simulations that the adaptive approach can reduce the expected sample size while keeping the power of the procedure and thus it can improve with respect to standard approaches which are based on a single test statistic. We have also employed the currently fashionable RMST to examine the difference between the two survival curves.

A series of points apply for the developed design as well as points for further investigation and improvement. The adaptation that we proposed is based on the formula 3.16 for estimating the sample size that we need. This formula uses the variance of the estimated RMST difference at the point of the estimation. And the point of the estimation in the adaptive design is when the “first sample” has been enlisted. It is clear that the better estimation of the variance we have, the better will be the estimation of the sample size. With the current design we estimate the variance based on a small sample and with an extrapolation because we estimate the variance of the RMST difference at a future t^* value. So an idea for further investigation is to construct a design that achieves better estimation for the variance. For that purpose we could use the data maturity analysis that is mentioned at Section 4.4 and to apply the formula 4.17.

The data maturity analysis shows when the sample size is adequate for the testing. So it gives us the opportunity to estimate the maturity at different stages of the accrual period and to stop enlisting patients if the maturity has been reached. This process provides us better estimations of the variance and as a result more accurate estimation for the sample size that we need. The adaptive design that we proposed can be modified to include the data maturity analysis. The first step of the design will be almost the same. The initial sample size will be divided into “first sample” and “second sample”. However the “second sample” will be also divided into smaller equal parts.

When the “first sample” has been enlisted then again we will test the proportionality assumption and if it is not rejected we will continue as described in the adaptive design. However, if the assumption is rejected we will not apply the formula 4.16 that estimates the sample size needed based on the RMST difference test. We will apply the data maturity analysis and the formula 4.17. If the data are estimated mature, then we can stop enlisting patients and go immediately to the follow up period. If the data are estimated to not be mature, then we will enlist one of the smaller parts of the “second sample” and then we will estimate again the data maturity analysis. We will continue this procedure even when the “second sample” has been enlisted and we will not stop until we reach the maximum sample size. If the maturity has not been reached until then, we will stop enlisting more patients and we will go to the follow up period.

The problem with this design that needs to be investigated is the problem of multi-testing. Since we estimate the data maturity several times, we increase the probability to estimate the data mature when they are not. More formally, we increase the probability of Type I error. So the further investigation should be in an adaptive design that uses the data maturity analysis, in order to have better estimation of the sample size, but also secures that we are not increasing the Type I error.

As already mentioned the current procedure is based on the RMST. Since RMST is less examined and used in clinical practice one can use as an adaptation any other test that avoids the proportionality assumption like the log-rank test. Such a test can be the Fleming-Harrington test statistic (with selected weight). We have not exploited this in the current thesis but it would be useful to see the performance for such an adaptation.

Finally as Grant et al (2014) have shown the test for proportionality assumption is a crucial choice. The one employed in this thesis by Grambsch and Therneau does not have adequate power in certain scenarios and perhaps the test is not able to discriminate between proportionality and non-proportionality. This can have an effect on our derived results. It would be useful to examine other tests of proportionality in order to improve the gain from the adaptive procedure discussed here.

References

1) Andersen PK, Perme MP (2010). Pseudo-observations in survival analysis, *Statistical Methods in Medical Research*;19(1):71–99
2) Chow S.-C., Chang M. (2008). Adaptive design methods in clinical trials – a review, *Orphanet Journal of Rare Diseases* 3:11
3) Cox, D.R. (1972): Regression models and life-tables, J. *R. Stat. Soc. Ser. B Methodol*. 187–220
4) David W. Hosmer, Stanley Lemeshow, Sussane May (2008). *APPLIED SURVIVAL ANALYSIS. Regression modeling of time-to-event data.* (2nd edition). John Wiley & Sons Inc., Hoboken, New Jersey.
5) Dirk F. Moore (2016). *Applied Survival Analysis Using R*. (1st edition). Springer International Publishing Switzerland
6) Freedman, L.S. (1982). Tables of the number of patients required in clinical trials using the log-ranktest, *Statistics in Medicine*. 1: 121-129
7) Grambsch P. and Therneau T. (1994), Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika*, 81, 515-26.
8) Grant, S., Chen, Y. Q., & May, S. (2014). Performance of Goodness-of-Fit Tests for the Cox Proportional Hazards Model with Time-Varying Covariates, *Lifetime Data Analysis*, 20(3), 355–368
9) Hajime Uno, Brian Claggett, Lu Tian, Eisuke Inoue, Paul Gallo, Toshio Miyata, Deborah Schrag, Masahiro Takeuchi, Yoshiaki Uyama, Lihui Zhao, Hicham Skali, Scott Solomon, Susanna Jacobus, Michael Hughes, Milton Packer, and Lee-Jen Wei (2014). Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis, *J Clin Oncol* 32:2380-2385
10) Harrington, D. P., Fleming, T. R. (1982). A class of rank test procedures for censored survival data, *Biometrika* 69, 553-566.
11) Huang B, Kuan P-F (2018). Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point, *Pharmaceutical Statistics*, 17:202–213.

12) Kaplan E. L. & Meier Paul (1958). Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association*, 53, 457-481

13) Mahajan Rajiv, Gupta Kapil (2010). Adaptive design clinical trials: Methodology, challenges and prospect, *Indian J Pharmacol*. 42(4): 201–207. doi: 10.4103/0253-7613.68417

14) Meinert, C. (1996). *Clinical trials dictionary: usage and recommendations*. Harbor Duvall Graphics, Baltimore, MD.

15) Rich, J. T., Neely, J. G., Paniello, R. C., Voelker C. C. J., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. Otolaryngology--Head and Neck Surgery, *Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, *143*(3), 331–336.

16) Rosner B. (2006). *Fundamentals of Biostatistics*. (6-th edition). Thomson Brooks/Cole.

17) Royston P, Parmar MKB (2011): The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt**,** *Stat Med* 2011, 30:2409–2421.

18) Royston P, Parmar MKB (2013): Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome, *Medical Research Methodology*,13:152

19) Schoenfeld DA (1983). Sample-size formula for the proportional-hazards regression model, *Biometrics*;39:499-503

20) Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model, *Biometrika* 69, 239-41.

21) Therneau T (2015). _A Package for Survival Analysis in S_. version2.38, <URL: https://CRAN.R-project.org/package=survival>.

22) Thomas D. Cook, David L. DeMets (2008). *Introduction to Statistical Methods for Clinical Trials*. Chapman and Hall/CRC, New York.

23) Wei, Y., Royston, P., Tierney, J. F., and Parmar, M. K. B. (2015) Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data, *Statist. Med.*, 34: 2881–2898.

24) Weiliang Qiu, Jorge Chavarro, Ross Lazarus, Bernard Rosner, Jing Ma (2018). Power and Sample Size Calculation for Survival Analysis of Epidemiological Studies. https://CRAN.R-project.org/package=powerSurvEpi

25) Xiaodong Luo (2017). PWEALL: Design and Monitoring of Survival Trials Accounting for Complex Situations. R package version 1.2.0. https://CRAN.R-project.org/package=PWEALL

26) Πανάγου Φ. (2014). Περιορισμένος Μέσος Χρόνος Επιβίωσης και Εφαρμογές στη Μετα – *Ανάλυση, Διπλωματική εργασία ΜΠΣ*, Αθήνα: Πανεπιστήμιο Αθηνών