Summer 8-9-2024

# The Impact of “Multiple Looks” when Performing Survival Analysis

Quentin Eloise
eloiseqy@jacks.sfasu.edu

# The Impact of “Multiple Looks” when Performing Survival Analysis

## Creative Commons License

# The Impact of "Multiple Looks" when Performing Survival Analysis

by

Quentin Eloise, B.S.

Presented to the Faculty of the Graduate School of

Stephen F. Austin State University

In Partial Fulfillment

of the Requirements

For the Degree of

Master of Science

STEPHEN F. AUSTIN STATE UNIVERSITY

August 2024

The Impact of "Multiple Looks" when Performing Survival Analysis

by

Quentin Eloise, B.S.

APPROVED:

Jacob Turner, Ph.D., Thesis Director

Derek Blankenship, Ph.D., Committee Member

Robert Henderson, Ph.D., Committee Member

Emiliano Giudici, Ph.D., Committee Member

Forrest Lane, Ph.D.
Dean of Research and Graduate Studies

## ABSTRACT

Survival analysis is a critical statistical method in healthcare to assess patient treatment effects and disease progression. Another critical area of statistical methodology in health care is the practice of adaptive designs. Adaptive designs allow for interim analyses to take place during a study and various decisions and actions can take place more ethically. This is beneficial for studies that take multiple years to complete and allows administrators and healthcare providers to make sound decisions as early as possible. A challenging aspect of adaptive designs is that the number of interim analyses is known in advance which is applicable in controlled experiments such as randomized clinical trials.

Motivated and highlighted by our collaborations with Fresenius Medical Care, many clinical studies are observational in nature and have no clear endpoint, making it difficult to determine the number of interim analyses that will be conducted. This research considers the application of survival analysis using adaptive designs within observational studies. To do so, we developed a collection of statistical programs to simulate these types of interim analyses while accounting for the additional complexity that survival data exhibits. Simulations summaries were performed and we will summarize some of the key results including investigations of statistical power, Type-I error control, and parameter estimation performance. Additionally, this work aims to assess the necessary conditions to achieve reasonable power at early looks and/or establish general rules of thumb when designing the study.

## ACKNOWLEDGEMENTS

Je souhaite tout d'abord exprimer ma profonde gratitude envers le Dr. Jacob Turner pour son soutien constant et ses conseils précieux tout au long de ce travail. Sa patience, son engagement et sa disponibilité ont été d'une importance capitale pour la réalisation de cette thèse.

Je tiens également à remercier le Dr. Derek Blankership, mon directeur de stage, pour son mentorat, sa gentillesse et sa positivité. Grâce à lui, mon expérience de stage a été très agréable et enrichissante.

Mes remerciements vont également au Dr. Robert Henderson et au Dr. Emiliano Giudici pour avoir accepté de consacrer leur temps à évaluer ce travail et pour leurs commentaires constructifs qui ont contribué à son amélioration.

Je souhaite exprimer ma reconnaissance envers mes camarades de classe pour leur soutien et leurs encouragements tout au long de cette aventure académique. Grâce à eux, j'ai passé deux très belles années ici à SFA.

Mes sincères remerciements vont à ma mère, mon père et ma sœur pour leur amour, leur soutien indéfectible et leur compréhension pendant cette période exigeante.

Enfin, je tiens à exprimer ma gratitude envers tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce travail.

## CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# 1 INTRODUCTION

Survival analysis is a general statistical technique for time-to-event outcomes. It is used in medical and healthcare research for numerous reasons, including the assessment of treatment effects and disease progression. Survival analysis can be applied to both observational and experimental study designs. In randomization prospective clinical trials, adaptive designs allow for interim analysis to take place during a study and various decisions and actions can take place more ethically. This is beneficial for studies that take multiple years to complete and allows administrators to make sound decisions as early as possible.

For this thesis, our interest is to effectively quantify the various statistical properties of providing an observational survival analysis in an adaptive design setting. Simulations will be conducted to get a better understanding of the advantages and potential shortfalls of conducting adaptive designs, as well as determining general rules of thumb or recommendations for best statistical practices.

In the introductory chapter, we will delve into the realm of survival analysis, followed by an exploration of epidemiology and adaptive designs. By doing so, we aim to elucidate the profound implications that adaptive designs can have on observational time-to-event studies. In the second chapter, we will go over the methodology employed to simulate survival data, ensuring it meets the Cox proportional hazard assumptions. We will outline the simulation function generated for this purpose. Subsequently, we will explore the function enabling "multiple look" data analysis once it's simulated or collected. In the third chapter, we will summarize the results of various simulations to observe if the input parameters influence the power of the test conducted at preceding interim analyses. We will also have a look at the overall Type

I error rate. The fourth and final chapter will summarize our findings and discuss potential directions for future research.

## 1.1 Introduction to Survival Analysis

To help introduce the key concepts and definitions of typical survival analysis, a data set curated by the North Central Cancer Treatment Group is utilized [7]. In this example, we are studying the time to death of patients with advanced-stage lung cancer. We have slightly modified the publicly available data set to illustrate some key logistical issues that arise in data processing. The table below provides the first 6 rows of the data set.

| | time | status | age | sex | ph.ecog | ph.karno | pat.karno |
|---|---|---|---|---|---|---|---|
| 1 | 306 | 1 | 74 | 1 | 1 | 90 | 100 |
| 2 | 455 | 1 | 68 | 1 | 0 | 90 | 90 |
| 3 | 1010 | 0 | 56 | 1 | 0 | 90 | 90 |
| 4 | 210 | 1 | 57 | 1 | 1 | 90 | 60 |
| 5 | 883 | 1 | 60 | 1 | 0 | 100 | 90 |
| 6 | 1022 | 0 | 74 | 1 | 1 | 50 | 80 |

Figure 1.1: First 6 rows of the lung cancer dataset

The variable "time" (in days), serves as our response variable and represents the survival time of the patients until either the event (death) happens or the data is censored. The survival time of an individual is said to be censored when the end point of interest has not been observed for that individual. This may be because of complete follow-up meaning the patient is still alive at the end of the study or because the patient has been lost to follow-up also called dropouts. In this example and the following chapters, we will focus only on right censoring. This censoring occurs when it is known that the event of interest occurred after a certain time $t$, that is to the

right of the last known survival time. In the data set used in this example, the categorical variable "status" takes the value 0 if the patient is censored or the value 1 if the patient is dead. The categorical variable "ph.ecog" represents the ECOG performance score as rated by the physician. 0=asymptomatic, 1 = symptomatic but completely ambulatory, 2 = in bed $<$ 50% of the day, 3 = in bed $>$ 50% of the day but not bedbound and 4 = bedbound. The "ph.karno" variable represents the Karnofsky performance score (bad=0-good=100) rated by the physician. The "pat.karno" variable represents the Karnofsky performance score as rated by patients, also from a scale of 0 to 100. The last two variables "meal.cal" (kcal) and "wt.loss" (pound) represent respectively the calories consumed at meals and the weight loss in the last six months. The survival time might also depend on these additional variables so models that can incorporate multiple explanatory variables can be very helpful.

In a typical prospective study, most patients are recruited simultaneously but accrue over months or even years. After recruitment, patients are followed until the outcome event occurs or until the end of the study time of the trial. In most studies, the recruitment period and follow-up period have the same length.

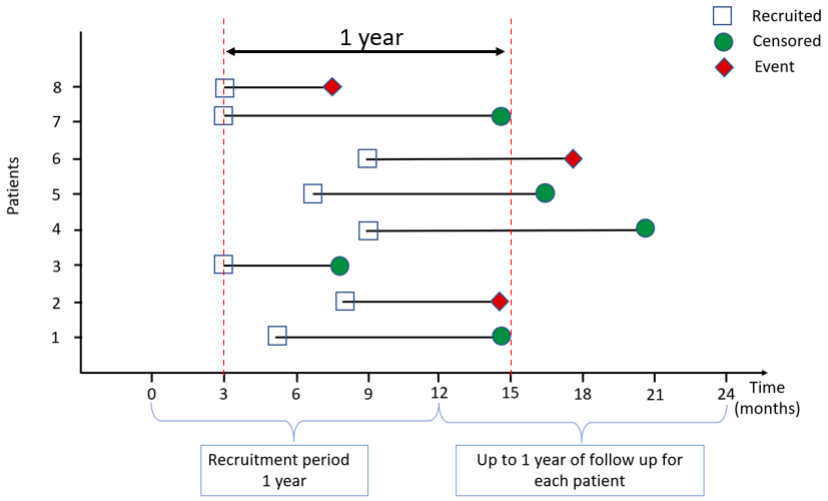


Figure 1.2: Example of patients within a study

In the example above, the recruitment period and subsequent follow-up for each participant extend over a year. Patients 3, 7, and 8 entered the study during the third month, signifying that they would be monitored until month 15, covering one year of follow-up. Patient 3 withdrew from the study in the ninth month, resulting in censorship due to drop-out after six months of the study. Patient 7, having not encountered the event by month 15, is censored due to the completion of the entire follow-up period. Conversely, Patient 8 experienced the event in the ninth month, indicating a time-to-event occurrence six months after enrollment in the study.

In analyzing survival data, three functions are of central interest: the survivor function, the hazard function, and the cumulative hazard function. The survivor function describes the probability of individuals surviving to or beyond a given time. Let the actual survival time of an individual be $t$, and let $T$ be the random variable associated with the survival time. Suppose that this random variable has a probability distribution with underlying probability density function $f(t)$. The distribution function of $T$ is then given by

$$F(t) = P(T \leq t) = \int_0^t f(u)du. \tag{1.1}$$

The function $F(t)$ represents the probability that the survival time is less than some value $t$. This function is called the cumulative incidence function because it summarizes the cumulative probability of an event occurring before time $t$. As previously defined, the survivor function $S(t)$ is the probability of individuals surviving to or beyond a given time t, so we have

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t). \tag{1.2}$$

The hazard function is used to express the risk or hazard, of an event occurring for an individual at some time $t$. The function $h(t)$ is obtained from the probability

that an individual dies at time $t$ conditional on the individual surviving to that time,

$$h(t) = \lim_{\delta t \to 0} \frac{Pr(t \leq T < t + \delta t | T \geq t)}{\delta t}. \tag{1.3}$$

Using the properties of conditional probabilities and the limit definition of a derivative, equation (1.3) can be expressed as

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.4}$$

The cumulative hazard function, H(t) is the cumulative risk of an event occurring by the time t. The cumulative hazard function at time t can also be interpreted as the expected number of events that occur in the interval from the time origin to t. From the equation (1.4) it follows that

$$h(t) = -\frac{d}{dt} log S(t). \tag{1.5}$$

and so

$$S(t) = e^{-H(t)}. \tag{1.6}$$

thus we have

$$H(t) = -log(S(t)). \tag{1.7}$$

For a more detailed discussion, see [4].

Survival analysis aims to achieve two main objectives. The first one is to estimate the probability of not experiencing an event of interest ("surviving") over any given time period. The second one is to compare the overall survival experience between different groups of individuals or determine if the survivor function is associated with covariates.

When comparing survival times between groups, an initial step is to present numerical or graphical summaries of the survival times for individuals in a particular

group. Survival data are usually summarised through estimates of the survivor function and hazard function. The methods used are non-parametric since they do not require specific assumptions about the distribution of the survival times.

Theoretically, the survivor function is a smooth curve, but for observed data, we have a finite number of subjects, so the estimate of the survivor function is a step function. The Kaplan-Meier estimate is the most commonly used method to estimate and visualize the survivor function. After arranging the times to death in ascending order, let time intervals $t_j$, where each $j$ indicates the ordered death time starting from 1, and $d_j$, the number of events that occurred in the interval $t_j$, and $n_j$, the number of individuals known to have not had an event occur or are right-censored up to time $t_j$ the Kaplan-Meier estimator is defined as

$$\hat{S}(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i}. \tag{1.8}$$

The Kaplan-Meier curve gives us an estimation of the probability that a subject survives longer than time $t$.

The Kaplan-Meier curve for the lung data set is displayed in Figure 1.3.

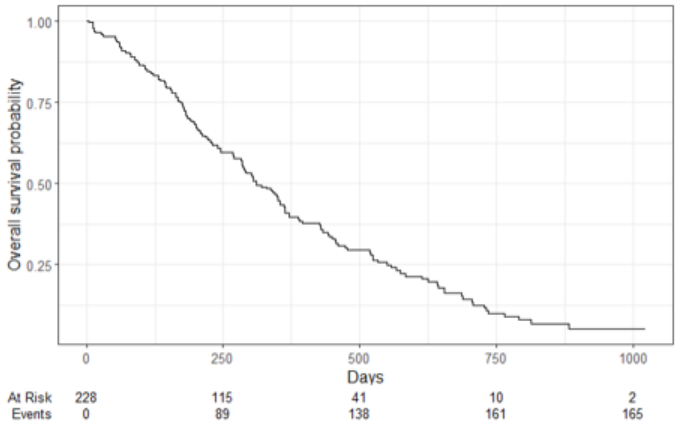


Figure 1.3: Kaplan-Meier Curve

As shown above, The Kaplan-Meier curve provides us with an easy way of visualizing the expected survival duration for patients over time. Through this curve, we can approximate the duration required for a specified percentage of patients to survive. In the context of the lung data set, the visualization indicates that approximately 50 percent of patients have survived at least 312 days. Conversely, we can deduce the percentage of patients who survived beyond a specific time frame. In the case of the lung data set, the curve suggests that only 27 percent of patients survived for more than 500 days.

A component of this thesis is to compare the overall survival between different groups of individuals. For the lung data set, we will look into the difference in the distribution of survival times between males (group 1) and females (group 2) using the Kaplan-Meier estimates.

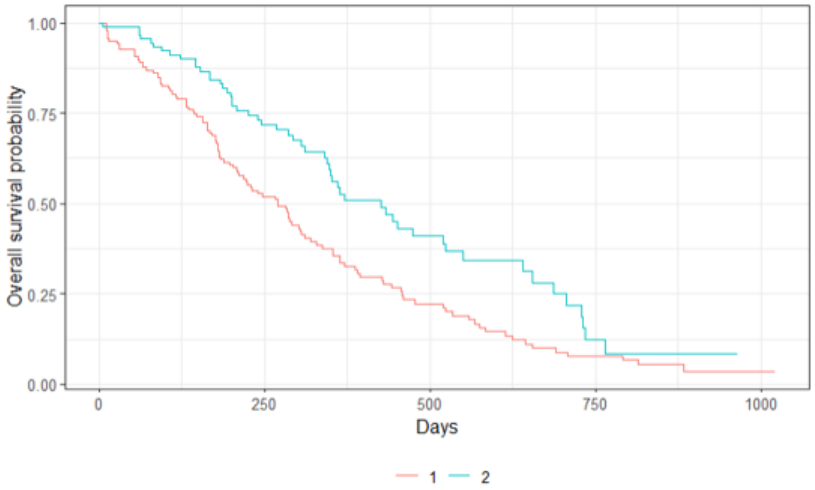


Figure 1.4: Kaplan-Meier Curves by sex

In Figure 1.4, a visual distinction is apparent. Group 2 (female) exhibits a superior survival probability over time in comparison to Group 1 (male). However, the Kaplan-Meier estimates alone do not offer information on the statistical significance of this difference or the magnitude of the disparity between the two groups.

The log-rank and Wilcoxon tests are commonly employed for comparing the survival functions of the two groups. The log-rank test is more powerful when the hazard of death at any given time for an individual in one group is proportional to the hazard at that time for a comparable individual in the other group. This assumption of proportional hazards is pivotal in various methods for survival data analysis. In instances where the assumption is not met, the Wilcoxon test is deemed more appropriate. While these methods can identify differences between groups, they do not provide an estimate of the size of the difference between them. [4]

The method that we will be focusing on to quantify the difference between groups is the Cox Proportional Hazards model. To use this model, the survival data should meet two assumptions. The first assumption is non-informative censoring also called random censoring which occurs when the patients drop out of the study for reasons

unrelated to the event being studied. The second assumption is proportional hazards. Let $h_0(t)$ be the hazard function for the control group and let $h(t)$ be the hazard function for the treatment group. If the two hazards are proportional, then we can say that

$$h(t) = \phi h_0(t), \tag{1.9}$$

where $\phi$ is a constant that does not depend on time. The proportional hazards assumption can be checked using statistical tests and graphical diagnostics such as Schoenfeld residuals. In general, the Schoenfeld residuals are independent of time. A plot that shows a non-random pattern against time is evidence of a violation of the non-proportional hazard assumption. In R, the function cox.zph() from the survival package provides a convenient solution to test the proportional hazard assumption for each variable in a Cox model. A non-significant relationship between residuals and time supports that the proportional hazard assumption is met.

Under the proportional hazard model assumption $h_c(t) = \phi h_t(t)$, $\phi$ is known as the hazard ratio. If $\phi > 1$, then we could say that the chance of survival is greater for the control group. If $\phi < 1$, we would say the opposite.

If we have more predictor variables, then the model can be extended similarly to that of multiple linear regression or logistic regression. The full Cox Proportional Hazards model is expressed by including linear predictors through $log(\phi)$,

$$ln(\phi) = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n. \tag{1.10}$$

Substituting equation (1.10) into equation (1.9), the Cox proportional hazard model is defined as

$$h(t) = (e^{\beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n}) h_0(t). \tag{1.11}$$

It should be noted that the previously defined hazard function for the control group, $h_0(t)$, is also called the baseline hazard in the general sense. The predictor variables could be either numeric or categorical. Unlike multiple linear regression, it should be

noted that the Cox Model does not have an intercept as any constant intercept term added to the model becomes integrated into the baseline hazard.

Another advantage of the Cox model is its interpretation of the regression coefficients. The model provides hazard ratios for each covariate, which are easy to interpret. A hazard ratio greater than 1 indicates an increased hazard (risk of event), less than 1 indicates a decreased hazard, and equal to 1 indicates no effect. Upon algebraic manipulation of equation (1.11), we have

$$ln\frac{h(t)}{h_0(t)} = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n. \tag{1.12}$$

The regression coefficients, $\beta_1, ..., \beta_n$, represent the expected change in the natural log of the hazard ratio for a one unit change in the corresponding predictor holding all other predictors constant. Using software, the hazard ratio across two groups can easily be estimated after fitting the proportional hazards model, as well as a confidence interval for this hazard ratio. Parameter estimation is conducted by maximum likelihood (MLE) and hypothesis testing and confidence intervals are conducted using the asymptotic properties of MLE's [4].

A Cox proportional hazard model was fit for the lung data set with "sex" as the only covariate (male = group 2 and female = group 1). The Kaplan Meyer curves along with assuming results of the fit are provided in Figure 1.5

The p-value of 0.001 suggests strong statistical significance of a difference between males and females. Moreover, the 95 % confidence interval indicates that the hazard rate for female patients is estimated to be between 42% and 82% of the hazard rate for male patients, with an alpha of 0.05. This observation is depicted in Figure 1.5, where the survival curve for females consistently surpasses that of males.

This analysis and discussion highlight that the Cox proportional hazard model can be utilized for a simple two-group analysis. It is an important reminder that the Cox model has the flexibility to include additional explanatory variables to allow for estimates of the hazard ratio to be adjusted for other potential confounding variables

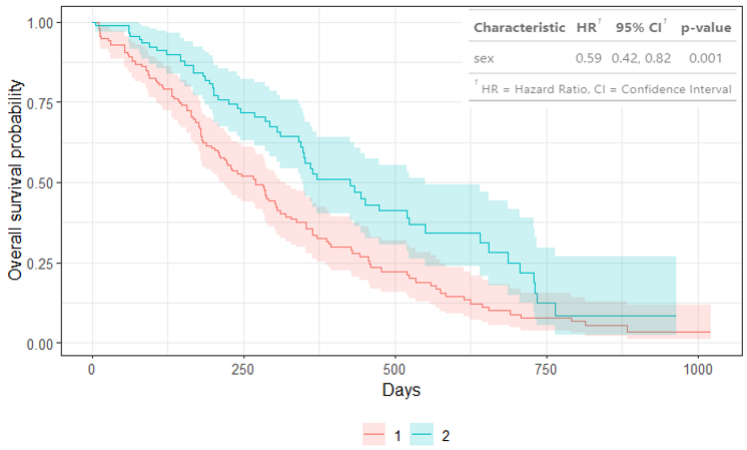


Figure 1.5: Kaplan-Meyer Curve by Sex with Cox Model estimates

in the model. For our purposes, we are mostly interested in the one predictor case, as we will be using propensity score matching to control for confounding factors across the two groups. This is standard practice for many researchers. For a more detailed explanation of propensity score matching see [12].

## 1.2 Introduction to Epidemiological and Adaptive Study Designs

In evidence-based research, the researcher should select the study design with the highest level of evidence possible as first described in a report by the Canadian Task Force on the Periodic Health Examination in 1979 [9]. The Canadian Task Force on the Periodic Health Examination was established in 1976 to determine how the Periodic Health Examination might enhance or protect the health of the population. The main goal was to recommend a plan for a lifetime program of periodic health assessments for all persons living in Canada [8]. As shown in table 1.1, the authors developed a rating system to determine the effectiveness of a particular intervention. The level of evidence was taken into account when evaluating a recommendation. For example, a Level 1 recommendation was given if there was good evidence to support

a recommendation.

Table 1.1: Adapted from Canadian Task Force on the Periodic Health Examination

| Level | Type of evidence |
|---|---|
| I | At least 1 RCT with proper randomization |
| II.1 | Well-designed cohort or case-control |
| II.2 | Time series comparisons or dramatic results from uncontrolled studies |
| III | Expert opinions |

Currently, in medical research a more detailed pyramid model illustrates the quality of available evidence [1].

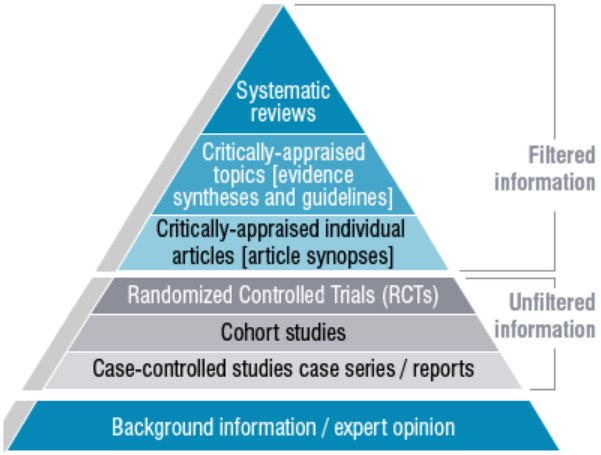


Figure 1.6: Levels of evidence pyramids

As shown in Figure 1.6 the higher the position on the pyramid the stronger the evidence. Levels of evidence pyramids are mostly divided into three sections. The top section groups the filtered evidence which contains synthesized information, such as systematic reviews and meta-analyses. This means that clinical experts and subject specialists have posed a question and then synthesized the available primary studies.

The second and third sections group the unfiltered evidence such as randomized control trials, observational studies, and, expert opinions.

When treating unfiltered information, well-conducted randomized controlled trials have been recognized as the gold standard for assessing the efficacy of clinical interventions, valued for their statistical rigor and mechanisms to avoid bias [3]. Next would be any evidence from cohort studies and case-control studies which are types of observational studies.

Observational studies are research designs in which researchers assess the association of some type of intervention, risk, or treatment without intervening or manipulating variables or subjects. These studies differ from experimental studies such as random control trials. When performing experimental studies researchers manipulate who is exposed to the treatment or intervention by having a control and treatment group [2]. A primary advantage to the observational study design is they can generally be completed quickly and inexpensively. Random clinical trials are more complex and involved, requiring many more logistics and details, whereas an observational study can be more easily designed and completed. Observational study designs also allow researchers to explore answers to questions where a randomized controlled trial would be unethical. The main disadvantage of observational studies is that they are more open to dispute than a randomized clinical trial. Observational studies are susceptible to confounding variables that can obscure true associations between variables and there is also a risk of biases, such as selection bias, that may affect the accuracy of the study's findings [2].

In 1965, British epidemiologist Sir Austin Bradford Hill proposed a set of nine principles, known as the Bradford Hill criteria, to evaluate the strength of evidence for a causal relationship between a specific factor and outcome in observational studies. These criteria encompass strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, and analogy. These criteria may lend

support for causality, but failing to meet some criteria does not necessarily provide evidence against causality either [5].

There are three types of observational studies based on sample/patient selection: case-control studies, cohort studies, and, cross-sectional studies. The data provided in our analysis comes from a cohort study, thus we will only describe that particular type of observational study. This type of observational study is often used to help understand cause and effect [2]. In a cohort study, a group of individuals with a common characteristic or exposure is followed over time, and the researchers observe and analyze the development of outcomes [6]. These studies involve comparing an exposed group to an unexposed group to ascertain whether the outcome of interest is associated to the exposure. Cohort studies are categorized into two types: prospective and retrospective. Prospective studies involve tracking a cohort into the future to observe health outcomes, whereas retrospective studies involve tracing patients back in time to gather cohort/exposure information and the occurrence of the outcome.[6].

While many prospective cohort studies are relatively brief, particularly those focusing on patients in the advanced stages of diseases like cancer, there are instances where such studies can extend over a longer duration. For instance, individuals with end-stage renal disease typically have a life expectancy ranging from 5 to 10 years. Consequently, studies within this domain often last over a considerable period. As with clinical trials, there is a natural inclination to want to perform preliminary or intermediate analyses, as well as a final analysis at the end point, and make decisions based on those results. This approach would be analogous to adaptive designs in clinical trials.

An adaptive design is defined as a design that allows modifications to a clinical trial and/or statistical procedures of the trial after its initiation without undermining its validity and integrity. The purpose is to make clinical trials more flexible, efficient, and fast [10]. Adaptive designs have been developed as an alternative to traditional

randomized controlled trial designs because traditional randomized controlled trials can demand substantial time and resources. In adaptive designs, patient outcomes are observed and analyzed at multiple predefined interim points called “looks” or “multiple looks”, and predetermined modifications to study design can be implemented based on the observations made from these multiple looks [10]. The next chapter will discuss the methodology used to simulate survival data and the implementation approach for the ”multiple looks.”

# 2 METHODOLOGY

In this chapter, we will delve into the methodology employed to simulate survival data, ensuring it meets the Cox proportional hazard assumptions. We will outline the simulation function generated for this purpose. Subsequently, we will explore the function enabling "multiple look" analysis of data once it's simulated or collected.

## 2.1 Simulation Function

In this thesis, we will simulate survival data that adhere to the Cox Proportional Hazards assumptions, without considering the need to address confounders through methods like propensity score matching. The study design replicates the one outlined in the TCU publication, spanning a two-year duration with patient recruitment occurring in the first year and patient follow-up lasting at least one year. Interim analyses are conducted every three months, totaling seven interim analyses, with a final analysis conducted at the study's conclusion. Thus, our data will consist of a time-to-event outcome and one binary covariate representing two treatment groups. Let X be an indicator variable that takes the value zero if an individual is in the control group and the value one if an individual is in the treatment group. Let $h_0(t)$ be the hazard function for the control group and let $h_1(t)$ be the hazard function for the treatment group. Assuming that the two hazards are proportional, as mentioned in the equation (1.9) we can say that

$$h_1(t) = \phi h_0(t),$$

If $x_i$ is the value of $X$ for the $ith$ individual in the study, i=1,2,...,n, the equation can be written for this individual as

$$h_i(t) = e^{\beta x_i} h_0(t),$$

This is the proportional hazards model for the comparison of two groups. Thus, to simulate survival data following the Cox Proportional Hazards assumptions, we would need to know the control group's hazard function and the $\beta$ coefficient. However, to generate realistic survival data, one must consider not only the distribution of survival times but also factors such as censoring and the duration of the study.

In general, to simulate data based on the outlined components, one would follow this framework:

1. Choose a distribution function, denoted as S(t), to model survival times.

2. Adjust the parameters of the chosen distribution to achieve the desired level of administrative censoring stemming from complete follow-up.

3. Appropriately select the $\beta$ coefficient based on the hazard ratio specified.

4. Address any remaining censoring resulting from dropouts and right censoring.

Based on this framework, a simulation function has been developed to facilitate the implementation of these steps. The simulation function has for input:

1. The hazard ratio ($\phi$).

2. The maximum length of the study (T).

3. The proportion of patients censored due to complete follow-up ($C_{cf}$).

4. The proportion of patients censored due to dropouts ($C_d$).

5. The number of control individuals.

6. The number of treatment individuals.

The output would be the survival data containing the categorical variable "$X$" that takes the value zero if an individual is in the control group and the value one if an individual is in the treatment group. The variable "time", serves as our response

variable and represents the survival time of the patients until either the event happens or the data is censored. The categorical variable "status" takes the value 0 if the patient is censored or 1 if the patient is dead. The selected distribution for the simulation function is the Weibull distribution. The Weibull distribution is a two-parameter distribution with shape parameter $\gamma$ and scale parameter $\lambda$ where $x \in [0, \infty)$, $\lambda > 0$ and $\gamma > 0$. For the following examples and results, we arbitrarily set $\gamma$=1 to obtain a Weibull distribution with a shape similar to the survival data analyzed by Fresenius Medical Care. The function was built respecting the following steps. First, an unconditional time to event (everyone dies) is generated following a Weibull distribution. Based on the latest time point during which observation may fail, we want the survivor function S(t) to equal the proportion of survived patients, as derived later in this thesis. In our simulation, the latest time point during which observation may fail is the maximum time of the study called T. Let $P(S)$ be the proportion of survived patients so we want

$$S(T) = P(S).$$

we know that $S(t) = 1 - F(t)$ from equation (1.2) where for the Weibull distribution, $\mathrm{F(t)} = 1 - e^{-(t/\lambda)^\gamma}$, and so we have

$$S(t) = 1 - (1 - e^{-(t/\lambda)^\gamma}) = e^{-(t/\lambda)^\gamma}.$$

we now have

$$S(T) = e^{-(T/\lambda)^\gamma} = P(S).$$

$$\gamma(\frac{-T}{\lambda}) = log(P(S)).$$

thus we have

$$\lambda = \frac{-T\gamma}{log(P(S))}.$$

Now that we have the parameter $\lambda$ required to generate the survival curve of the control group, we will now derive the $\beta$ coefficient to generate the survival curve of the treatment group as a function of the hazard ratio $\phi$. The hazard function is defined as

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

For the Weibull distribution, the probability density function is

$$f(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1} e^{-(t/\lambda)^{\gamma}}$$

and as previously derived the survivor function is

$$S(t) = e^{-(t/\lambda)^{\gamma}}.$$

Thus, the hazard function for the Weibull distribution is as follows:

$$h(t) = \frac{\frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1} e^{-(t/\lambda)^{\gamma}}}{e^{-(t/\lambda)^{\gamma}}}$$

$$h(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1}.$$

The hazard function for the control group (X=0) is:

$$h_0(t) = \frac{\gamma_0}{\lambda_0}\left(\frac{t}{\lambda_0}\right)^{\gamma_0-1}.$$

The hazard function for the treatment group (X=1) is:

$$h_1(t) = \frac{\gamma_1}{\lambda_1}\left(\frac{t}{\lambda_1}\right)^{\gamma_1-1}.$$

For the proportional hazard assumption to be met, we need the same shape of Weibull distribution in the treatment and control groups ($\gamma_0 = \gamma_1 = \gamma$). From equation (1.9) we have

$$h(t) = \phi h_0(t),$$

thus we have the hazard ratio $\phi$ which is

$$\phi = \frac{h_1(t)}{h_0(t)},$$

$$\phi = \frac{\frac{\gamma}{\lambda_1}\left(\frac{t}{\lambda_1}\right)^{\gamma-1}}{\frac{\gamma}{\lambda_0}\left(\frac{t}{\lambda_0}\right)^{\gamma-1}}$$

$$\phi = \frac{\left(\frac{1}{\lambda_1}\right)^{\gamma}}{\left(\frac{1}{\lambda_0}\right)^{\gamma}}$$

$$\phi = \left(\frac{\lambda_0}{\lambda_1}\right)^{\gamma}$$

As mentioned in Equation (1.11), $\lambda_0$ and $\lambda_1$ can be written as $\lambda_0 = e^{\beta_0}$ and $\lambda_1 = e^{\beta_0+\beta_1}$. Thus now we have

$$\phi = \left(\frac{e^{\beta_0}}{e^{\beta_0+\beta_1}}\right)^{\gamma}$$

Solving for $\beta_1$ yields

$$\beta_1 = \frac{-log(\phi)}{\gamma}.$$

Thus we have the required Beta coefficient to generate the treatment group survival curve from the user-specified hazard ratio $\phi$.

Once the treatment and control survival curves are generated, the observations with a time greater or equal to the maximum time will become censored (due to complete follow-up). Thus we have patients censored due to complete follow-up and patients that died (all patients with t<T). Initially, our simulated sample consists of both deceased and surviving patients (still alive) as depicted in Figure 2.1.

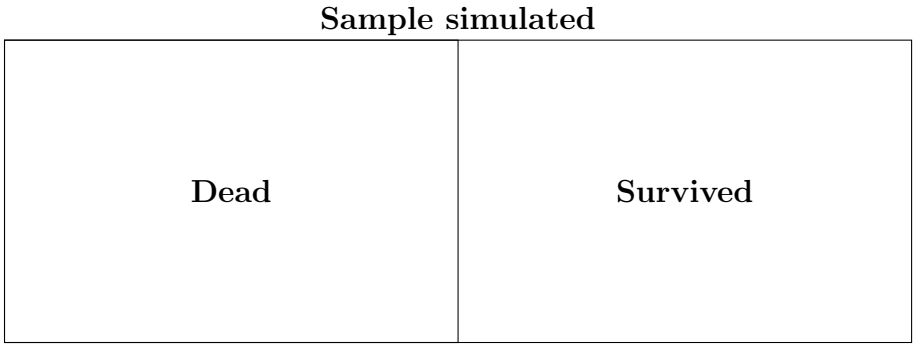


Figure 2.1: Partition space of simulated data

However, an additional form of censoring that requires consideration is censoring due to dropouts. These occurrences are assumed to be randomly distributed across our simulated sample. Figure 2.3 visually represents our updated simulated sample.

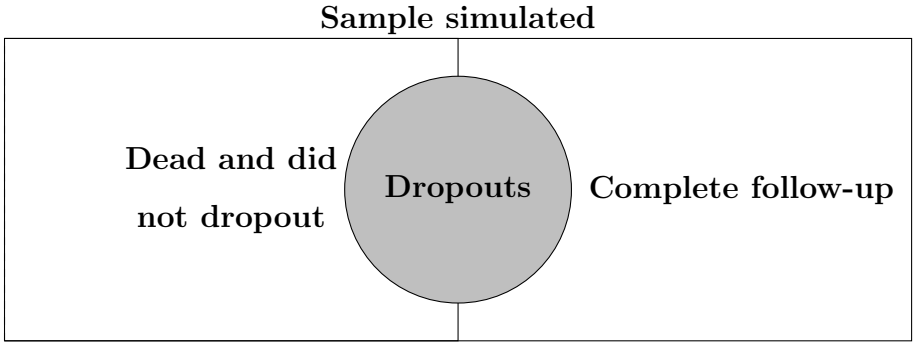


Figure 2.2: Updated partition space of simulated data

Consequently, a portion of the censoring attributed to dropouts originates from patients who survived, and the remaining patients who survived without dropping out are censored due to complete follow-up.

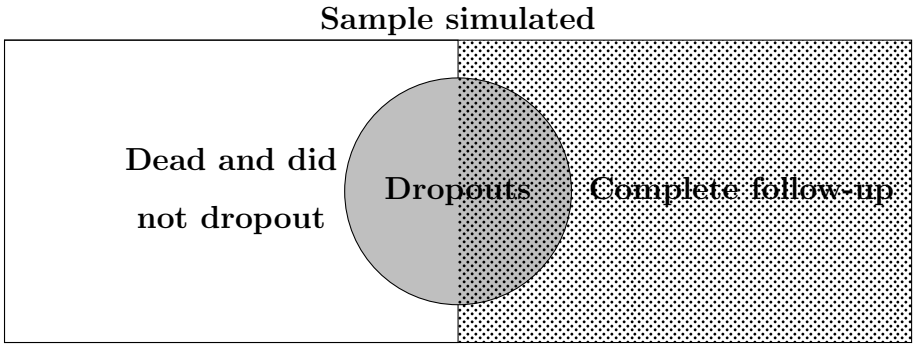


Figure 2.3: Updated partition space of simulated data

Let $C_{cf}$ represent observations censored due to complete follow-up and let $C_d$ represent observations censored due to dropouts. The proportion of patients who survived $=P(S)$, as shown in the shaded area in Figure 2.3 can be represented as:

$$P(S) = P(C_{cf}) + P(S \cap C_d).$$

As $S$ and $C_d$ are independent within the framework of our code, we can state that

$$P(S) = P(C_{cf}) + P(S) \cdot P(C_d)$$

$$P(S) - P(S) \cdot P(C_d) = P(C_{cf})$$

$$P(S)(1 - P(C_d)) = P(C_{cf})$$

$$P(S) = \frac{P(C_{cf})}{1 - P(C_d)}.$$

Thus, the proportion of patients who survived is derived from the provided proportions of those censored due to complete follow-up and dropouts. Letting $C_T$ denote

the total number of censored observations. In practice, an observation is either censored due to complete follow-up or due to dropouts, which are mutually exclusive events. Therefore, according to the probability addition rule, we have:

$$P(C_T) = P(C_{cf}) + P(C_d).$$

Using the above derivations and simulation strategy, a function has been developed to simulate survival data conforming to a Weibull distribution. This function allows us to manipulate parameters such as the hazard ratio, overall censoring, maximum study duration, and sample size. To ensure the coherence of our simulation function in generating survival data that conforms to Cox assumptions, we will simulate a dataset and juxtapose the theoretical curve with the Kaplan-Meier curve of the reference group. Additionally, we will conduct the Schoenfeld Test to confirm adherence to Cox proportional Hazards assumptions, estimate the Hazard ratio, and ascertain if the true Hazard ratio falls within a 95% confidence interval of the estimation.

## 2.2 Validating Simulated Data

To validate our simulation function, we will generate two datasets, compare their theoretical curves and Kaplan-Meir curves of the reference group, assess the Schoenfeld Test to verify that the Cox proportional Hazards assumptions are met, and evaluate their estimated and true hazard ratios. As previously mentioned we will set $\gamma$=1 for the following examples.

The initial dataset we simulated comprises 5,000 treatment and 5,000 control patients, with a maximum study duration of 365 days and a true hazard ratio of 0.75 and is depicted in Figure 2.4. The dashed blue line represents the theoretical curve S(t) and the grey solid line depicts the Kaplan-Meier curve estimate of the reference group. As we can see in Figure 2.4 the theoretical curve and the Kaplan-

Meier curve closely resemble each other, suggesting that our data aligns with the simulated expectations.

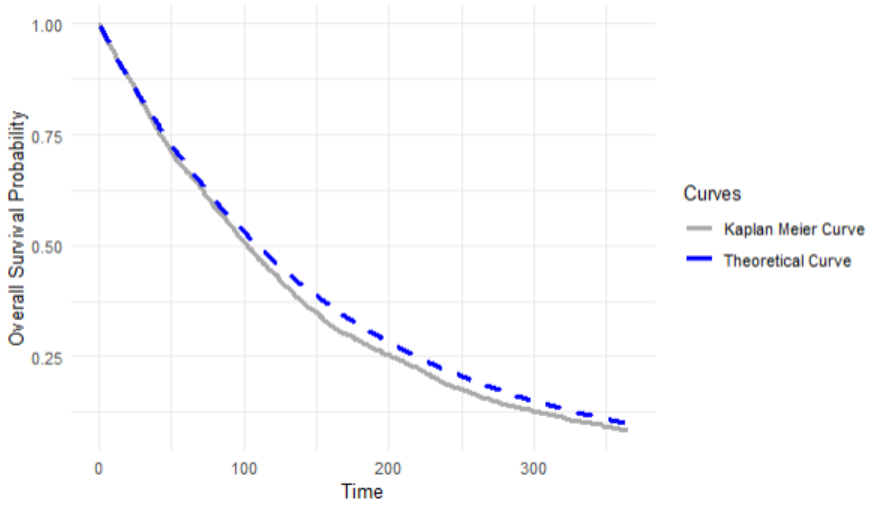


Figure 2.4: Theoretical and Kaplan Meier Survival Curves

With the simulated survival data closely resembling the expected theoretical curve, our next step involves verifying whether the proportional hazards assumptions are met. These assumptions can be assessed through statistical and graphical diagnostics, primarily focusing on the scaled Schoenfeld residuals using the survival library's cox.zph() function from R Studio.

In Figure 2.5, the solid line is a smoothing spline fit to the plot, with the dashed lines representing a $\pm\ 2\sigma$ band around the fit. The obtained p-value of 0.81 suggests non-significance, indicating that we find the proportional hazards assumption to be reasonable. Furthermore, upon visual inspection, the zero slope of the smoothing spline fit and all the residuals being within $2\sigma$ further elevates our confidence in the reasonableness of the proportional hazards assumption.

Having established that our simulated survival data adheres to the Cox proportional hazards assumptions, our next step involves examining the estimated hazard ratio. As depicted in Table 2.1, we observe that not only does the true hazard ratio

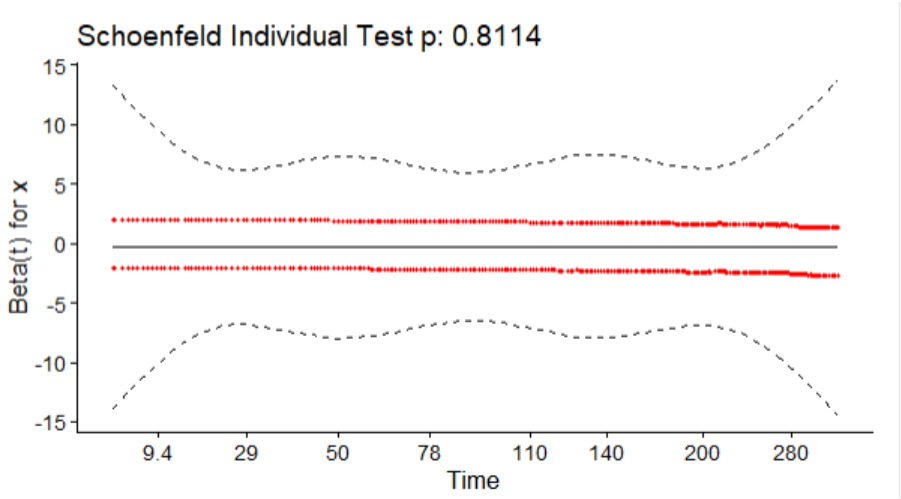


Figure 2.5: Schoenfeld residuals

fall within the 95% confidence interval (CI) of the estimated hazard ratio, but that the estimation and true value are identical.

Table 2.1: True and estimated Hazard ratio

| True Hazard Ratio | Estimated Hazard ratio | Lower 95% CI | Upper 95% CI |
|---|---|---|---|
| 0.75 | 0.75 | 0.69 | 0.83 |

For the second dataset, we will generate a comparatively smaller sample containing 100 treatment and control patients, with a maximum study duration of 365 days and a true hazard ratio of 0.5. As we can see in Figure 2.6 the theoretical curve and the Kaplan-Meier curve closely resemble each other, suggesting that our data aligns with the simulated expectations.

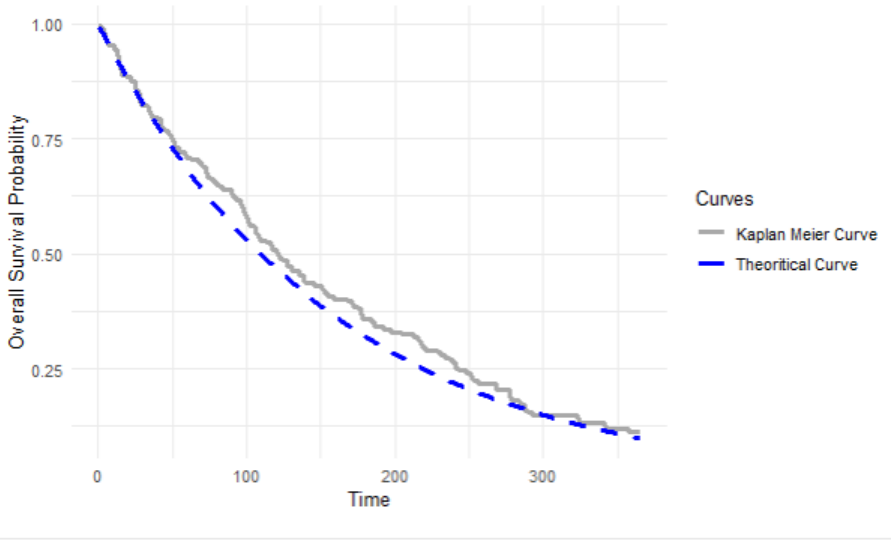


Figure 2.6: Theoretical and Kaplan Meier Survival Curves

With the simulated survival data closely resembling the expected theoretical curve, our next step involves verifying whether the proportional hazards assumptions are met using the scaled Schoenfeld residuals. In Figure 2.7, the obtained p-value of 0.66 suggests non-significance, indicating that we find the proportional hazards assumption to be reasonable. Furthermore, upon visual inspection, the zero slope of the smoothing spline fit and all the residuals being within $2\sigma$ further elevates our confidence in the reasonableness of the proportional hazards assumption.

Having established that our simulated survival data adheres to the Cox proportional hazards assumptions, our next step involves examining the estimated hazard ratio. As shown in Table 2.2, we note that not only does the true hazard ratio lie within the 95% confidence interval (CI) of the estimated hazard ratio, but also the estimation closely aligns with the true value.

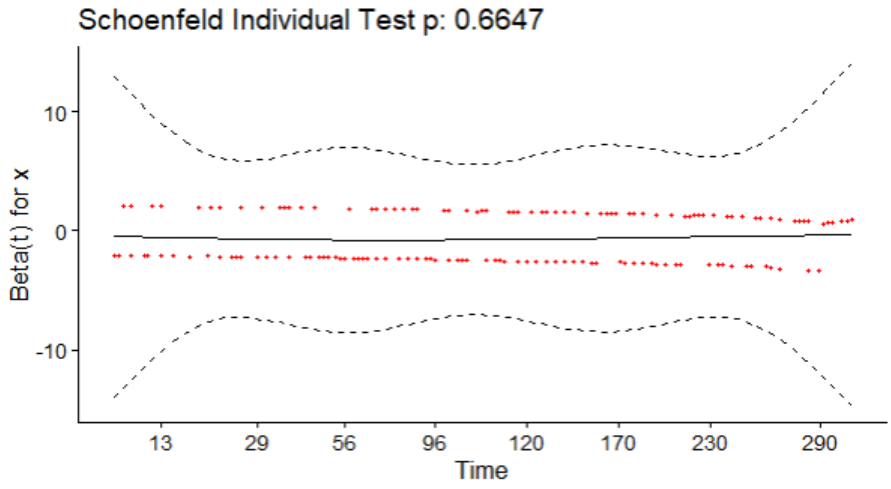


Figure 2.7: Schoenfeld residuals

Table 2.2: True and Estimated Hazard Ratio

| True Hazard Ratio | Estimated Hazard Ratio | Lower 95% CI | Upper 95% CI |
|---|---|---|---|
| 0.5 | 0.54 | 0.39 | 0.74 |

The analysis of these two datasets demonstrates the proficiency of the "Simulation function" in accurately simulating survival data, even when dealing with smaller datasets.

## 2.3 Multiple-look function

Projects involving survival analysis of individuals with end-stage renal disease, as discussed in Chapter 1, tend to be conducted over numerous years, and follow-ups are conducted throughout the studies and not just at the final endpoint. Similar to clinical trials, it is natural to want to perform preliminary analysis (looks) at an earlier time point or at multiple time points.

Now that we are capable of simulating survival data following a Weibull distri-

bution meeting the Cox proportional hazards assumptions, our next goal involves conducting multiple preliminary analyses on a simulated dataset. To facilitate this process, we have developed a "Multiple looks" function. This function requires the entire dataset under study as input, along with the desired number of looks to be performed. Upon execution, the function generates a summary table presenting the results obtained at each preliminary analysis. To align with end-stage renal disease studies, let the recruitment period be one year (365 days). Consequently, the range of starting dates for the simulated studies will span one year and will follow a uniform distribution. Additionally, let the follow-up period also be one year. The next step entails partitioning the input data into four equally spaced time intervals, each encompassing the preceding datasets, as illustrated in Figure 2.8. Since our simulation function generates the complete study dataset, for each preliminary analysis, it is imperative to update not only the patients' censoring status due to the differing endpoint but also their vital status (alive or deceased) may change. Consequently, we will examine the available data at intervals of six months, one year, one and a half years, and two years (the entire study duration).

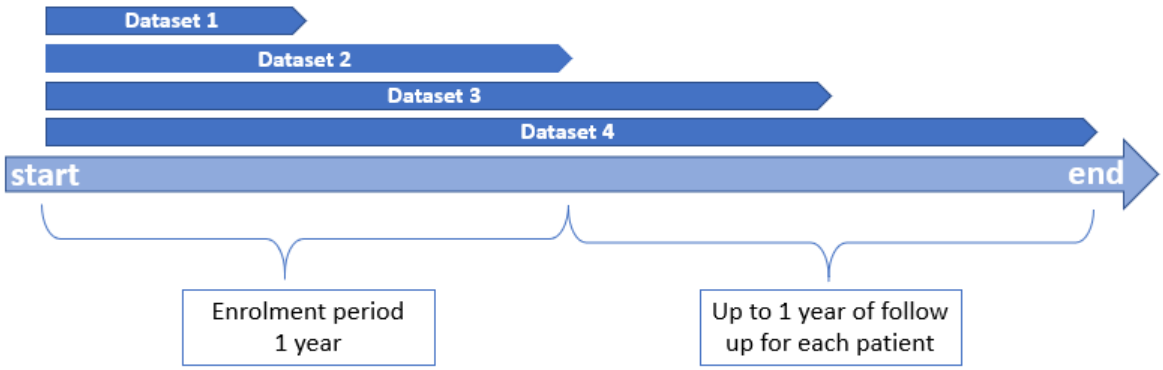


Figure 2.8: Multiple Looks visual

From each of the sub-datasets produced, the "Multiple look" function generates a result table. This table includes, for each look, the count of patients, the count of

deceased patients, the true and estimated hazard ratios, a 95% confidence interval for the estimated hazard ratio, and the associated p-value, as depicted in Figure 2.9.

| | look | observations | dead | h_low | hazard_Ratio | h_high | true_Hazard_Ratio | p_value |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1810 | 58 | 0.1968048 | 0.4582005 | 1.066782 | 0.913 | 0.07028532 |
| 2 | 2 | 3605 | 218 | 0.5280129 | 0.7580792 | 1.088390 | 0.913 | 0.13336019 |
| 3 | 3 | 3616 | 361 | 0.6732344 | 0.8791763 | 1.148116 | 0.913 | 0.34432242 |
| 4 | 4 | 3616 | 409 | 0.7192218 | 0.9204563 | 1.177995 | 0.913 | 0.51020591 |

Figure 2.9: Result table

With the capability to not only simulate survival data but also conduct analyses at multiple interim points, a pertinent question emerges: what is the power of the test conducted at each interim analysis, and do the input parameters exert any influence on the power of the test conducted at preceding interim analyses? The next chapter will address these questions through simulations of various scenarios.

# 3 SIMULATION RESULTS

In this chapter, we will summarize the results of various simulations to examine how input parameters influence the test's power during preceding interim analyses. Additionally, we will evaluate how these parameters affect the overall Type I error rate.

The relevant questions are: 1) What is the power of the test conducted at each interim analysis? 2) Do the input parameters exert any influence on the power of the test conducted at preceding interim analyses? To answer these questions, a comprehensive report covering the analysis at multiple interim points was assembled from 1,000 simulations for every 140 combinations of all the values assigned to each parameter under consideration, as presented in Table 3.1.

## 3.1 Simulation Results for Power estimation

Table 3.1: Simulation Parameters

| Parameter | Values |
|---|---|
| Hazard Ratio | 0.1, 0.25, 0.333, 0.5, 0.666, 0.75, 0.9 |
| Number of Treatment and Control | 100, 500, 1000, 5000 |
| Overall Censoring Rate | 0.1, 0.2, 0.5, 0.8, 0.9 |
| Recruitment Period | 1 year |
| Maximum Study Length | 1 year |
| Number of Interim Analyses/Looks | 8 |

The report includes a plot and a table summarizing the 1,000 simulations con-

ducted for each studied combination. The plot displays boxplots of the hazard ratios at each interim analysis across the thousand simulations. In the plot, the green dotted line denotes the true hazard ratio, while the red dot signifies the mean of the estimated hazard ratios. Additionally, the red whiskers represent a two-standard deviation interval. This plot provides an insight into the accuracy of our hazard ratio estimates at each interim analysis. Figure 3.1 displays one of the plots generated in the report. This plot was created for a simulation featuring a hazard ratio of 0.1, with maximum study and recruitment durations set to 365 days, and involving 500 control and treatment patients with an overall censoring proportion of 0.1.

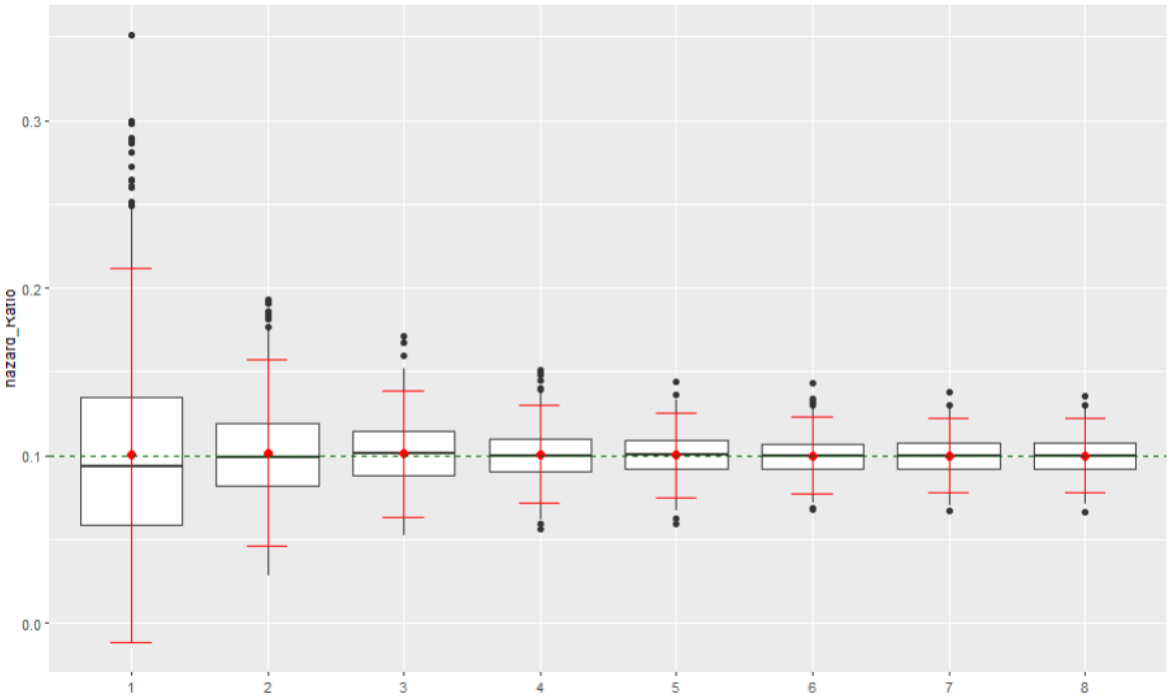


Figure 3.1: Report Boxplot Example

As shown in table 3.2, for each interim analysis (Look), the count of simulations (Sim), the average count of control (ACO) and treatment patients (ATO), and the average count of deceased control (ADC) and treatment patients (ADT). The final four columns indicate the proportion of instances where the test for a difference be-

Table 3.2: Report Table Example

| Look | Sim | ACO | ATO | ADC | ADT | PSP | PSN | PNSP | PNSN |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1000 | 123 | 127 | 33.7 | 4.0 | 0 | 0.978 | 0 | 0.022 |
| 2 | 1000 | 256 | 243 | 109.9 | 15.9 | 0 | 1 | 0 | 0 |
| 3 | 1000 | 382 | 366 | 206.9 | 34.7 | 0 | 1 | 0 | 0 |
| 4 | 1000 | 500 | 500 | 315.4 | 60.2 | 0 | 1 | 0 | 0 |
| 5 | 1000 | 500 | 500 | 395.7 | 85.0 | 0 | 1 | 0 | 0 |
| 6 | 1000 | 500 | 500 | 431.8 | 102.0 | 0 | 1 | 0 | 0 |
| 7 | 1000 | 500 | 500 | 446.2 | 111.9 | 0 | 1 | 0 | 0 |
| 8 | 1000 | 500 | 500 | 449.2 | 115.2 | 0 | 1 | 0 | 0 |

tween the control and treatment groups was deemed significant or not. However, we also aimed to indicate in table 3.2 whether the significance stemmed from the treatment or control groups. Thus, if the estimated hazard ratio is smaller or equal to 1, indicating that the treatment group outperforms the control group (with the treatment as the reference group), we categorized that significance as negative significance (PSN). Conversely, if the estimated hazard ratio is greater than 1, indicating that the control group outperforms the treatment group (with the treatment as the reference group), we categorized that significance as positive significance (PSP). Similarly, the non-significance proportion was divided into PNSP and PNSN. Figure 3.1 illustrates one of the tables outputted in the report. The presented table adheres to the same simulation parameters as the preceding plot.

Power is the chance of rejecting $H_0$ when $H_0$ is truly false. Put simply, power is the probability of rejecting when you are supposed to. However, within the context of this problem, it's established that all hazard ratios simulated will be less than or equal to 1. Consequently, we introduce a modified concept termed "probability of being correct," denoted power$^*$. This signifies the likelihood of rejecting the null

hypothesis in the correct direction, which means not only rejecting the null but also ensuring that the estimated hazard ratio is less than 1, note that power* is PSN. Hence, we will exclude the probability of rejecting in the incorrect direction, which entails disregarding instances where we reject the null hypothesis and the hazard ratio is greater than 1. It is worth noting that this estimated probability is approximately 0 in all cases.

The report groups 140 plots and tables. The following node charts summarize the information gained from the plots and tables. The metric that we chose to focus on is the "probability of being correct," also called power*. In the following node charts, "HR" denotes the Hazard Ratio, "N" signifies the count of treatment and control patients, and "C" indicates the overall censoring. The terminal nodes provide insights into the point at which the increase in power* ceases to be significant. Specifically, we define significance as any increase exceeding 5% from the initially identified look to the final one.

Our analysis produced numerous results that are applicable across the majority of parameter combinations observed in the report. Power* tends to rise alongside both sample size and the look number. Conversely, power* declines with higher overall censoring and as the Hazard Ratio approaches 1. After the fifth observation, which marks the conclusion of patient enrollment in the study, there is typically a minimal increase in power*. Next, the node plots will provide us with additional case-specific observations.

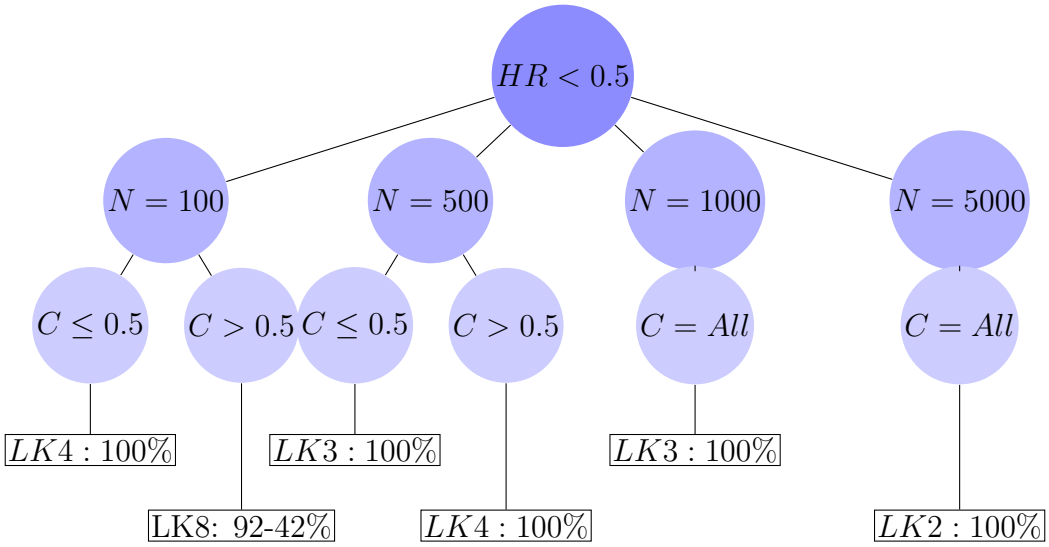


Figure 3.2: Summary for Hazard ratio $< 0.5$

When examining hazard ratios, a hazard ratio of 1 indicates no difference between the two groups, while values closer to 0 or $\infty$ signify a greater difference. A hazard ratio smaller than 1 indicates that the reference group has a higher survival rate compared to the other group. Conversely, a hazard ratio greater than 1 implies that the reference group has a lower survival rate compared to the other group.

Figure 3.2 presents a summary of findings concerning Hazard Ratios smaller than 0.5. It is apparent that in almost all instances where Hazard Ratios are below 0.5, there is no discernible increase in power* beyond the fourth look, as it maxed out at 100%. This indicates that early decisions with remarkably high power* can be confidently made when the survival rate in the treatment group exceeds double that of the control group.

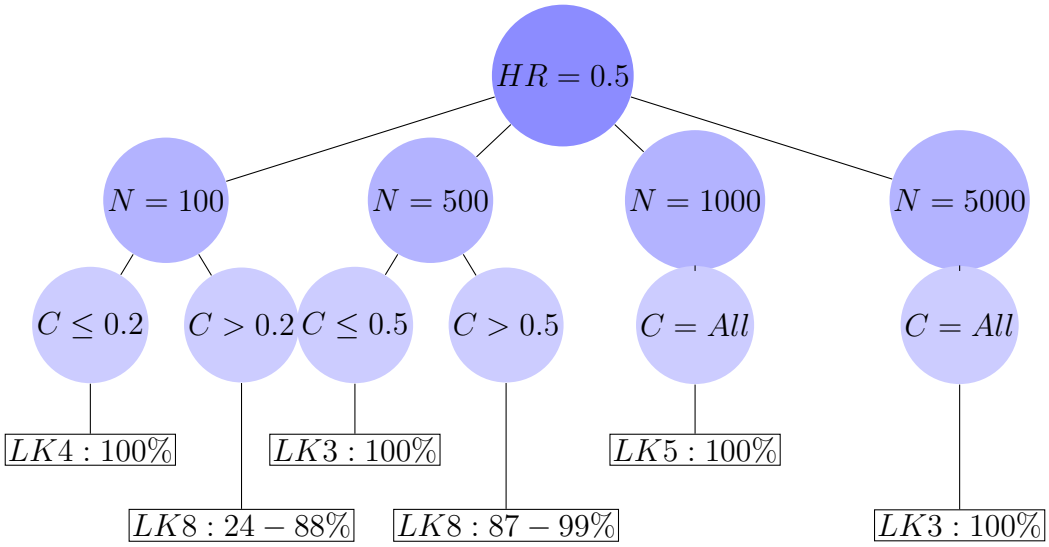


Figure 3.3: Summary for Hazard ratio = 0.5

Figure 3.3 provides a summary of findings regarding Hazard Ratios equal to 0.5. For sample sizes exceeding 1,000 in both treatment and control groups, power* reaches 100% after the fifth look. When the sample size is 500 and censoring is below 0.5, or when the sample size is 100 and overall censoring is below 0.2, power* reaches 100% after the fourth look. However, if the sample size is 500 with overall censoring exceeding 0.5, or if the sample size is 100 with overall censoring above 0.2, power* increases steadily until the final look.

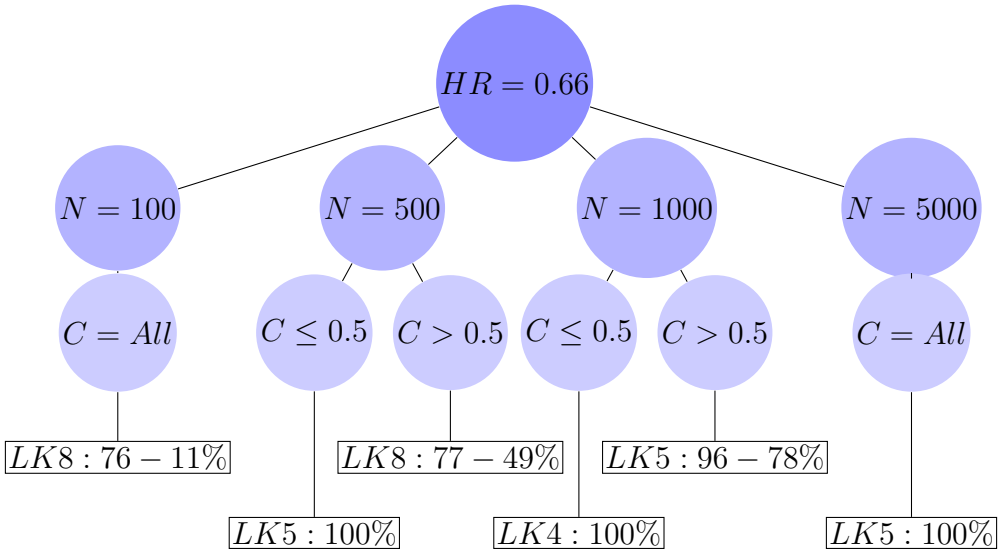


Figure 3.4: Summary for Hazard ratio = 0.66

Figure 3.4 summarizes the findings for Hazard Ratios equal to 0.66. When sample sizes exceed 5,000 in both treatment and control groups, or when overall censoring is below 0.5 for sample sizes of 500 and 1,000, power* reaches 100% after the fifth look. In all other cases, power* increases steadily until the final look.

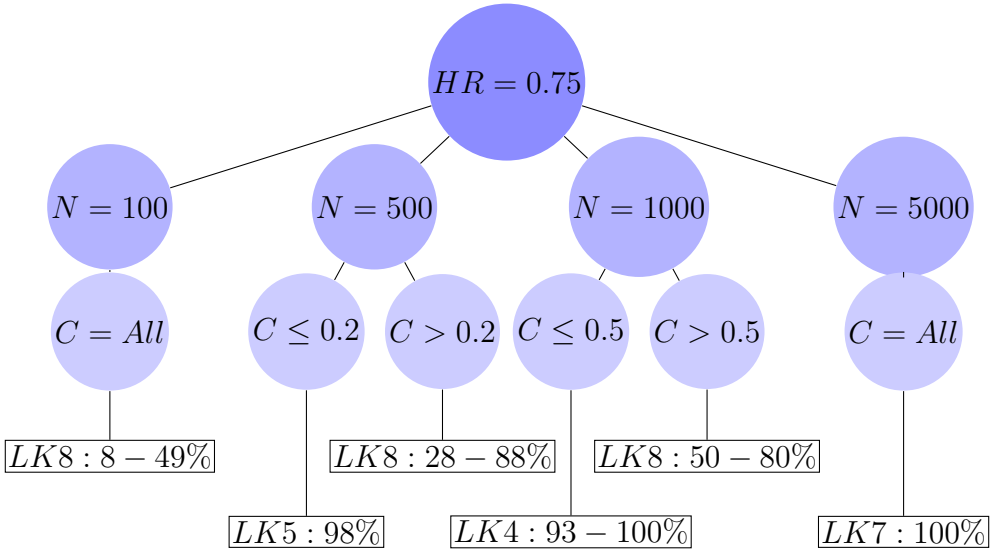


Figure 3.5: Summary for Hazard ratio = 0.75

Figure 3.5 summarizes findings regarding Hazard Ratios equal to 0.75. Power* achieves 100% only when the sample size is 5,000. However, for sample sizes of 500 and 1,000, power* does not surpass 98% after the fifth observation when censoring is below 0.2 and 0.5, respectively. In all other scenarios, power* steadily increases until the final observation.

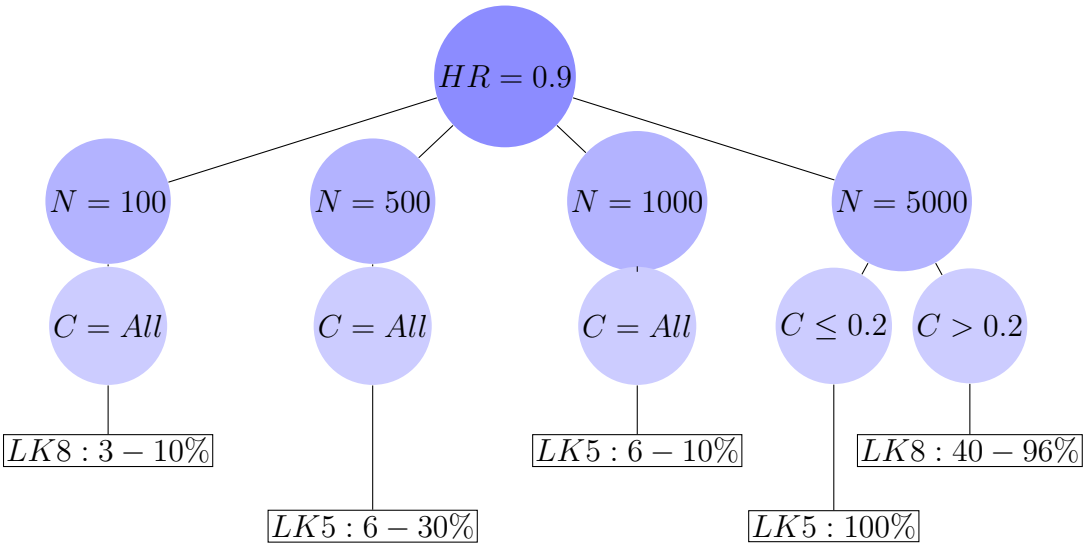


Figure 3.6: Summary for Hazard ratio = 0.9

Figure 3.6 summarizes findings regarding Hazard Ratios equal to 0.9. Power* reaches 100% only when the sample size is 5,000 and the censoring is below 0.2. However, for sample sizes of 500 and 1,000, power* does not increase beyond the fifth look.

From these node charts, the following summary emerges: as anticipated, power* increases with sample size, the number of observations, and the magnitude of the hazard ratio (diverging further from 1). Expectedly, power* declines with increased overall censoring. However, an unexpected finding is that after the fifth observation, denoting the conclusion of patient enrollment in the study, there is typically only a marginal increase in power*. Looking forward, these node charts provide valuable

guidance for future research, especially concerning including additional case-specific observations if required.

### 3.2 Simulation results for Type I error rate estimation

Having explored and summarized the information from the comprehensive report generated on the power at each interim analysis, we noticed the influence of the overall censoring rate and sample size on the power thus a pertinent question arises: How do these factors affect the Type I error at each interim analysis? The Type I error rate ($\alpha$) is defined as the probability of rejecting $H_0$ when $H_0$ is actually true. In our simulation setting, $H_0$ is considered true when simulating data with a true hazard ratio of 1. Hence, in order to evaluate the Type I error rate at each interim analysis in our simulation, we will evaluate the proportion of cases where the estimated hazard ratio was found to be statistically distinct from 1. The Type I error rate results are expected to align with the chosen significance level of 0.05 closely.

A comprehensive report covering the analysis at multiple interim points was assembled from 1,000 simulations for each combination of all the values assigned to each parameter under consideration, as presented in Table 3.3.

Table 3.3: Simulation Parameters

| Parameter | Values |
|---|---|
| Hazard Ratio | 1 |
| Number of Treatment and Control | 100, 500, 1000, 5000 |
| Overall Censoring Rate | 0.1, 0.5, 0.9 |
| Recruitment Period | 1 year |
| Maximum Length of the Study | 1 year |
| Number of Interim Analysis/look | 8 |

The subsequent graphs depict the mean Type I error across our 1,000 simulations, with simulation coverage and simulation error bar observed at each look. Let $\hat{\alpha}$ be the estimate of the Type I error rate. We define the simulation error bar as $\hat{\alpha} \pm 1.96\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{n}}$. Each graph corresponds to a specific sample size and all three censoring scenarios.

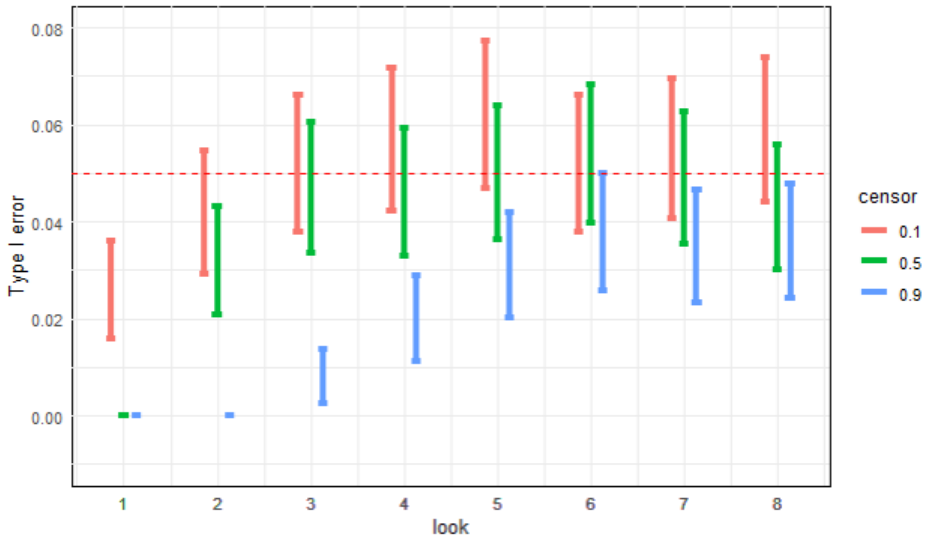


Figure 3.7: Type I error rate ± simulation error for HR=1 and sample size 100

In Figure 3.7, it is evident that with 100 control and 100 treatment patients, the Type I error rate aligns with the anticipated significance level from the third look for censoring rates of 0.1 and 0.5. However, for a censoring rate of 0.9, the Type I error consistently falls well below the expected significance level. Some of the rates close to 0 have a very tight interval due to the lack of number of rejections. It is known that the simulation error formula used can be liberal when the coverage is close to 0. Another simulation coverage based on Agresti-Coul or Bayesian formulas would be interesting to investigate.

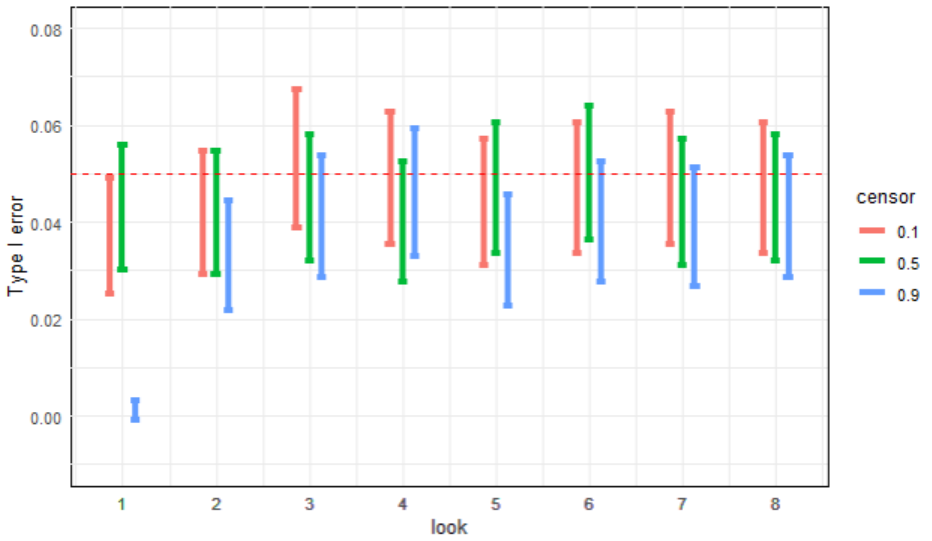


Figure 3.8: Type I error rate ± simulation error for HR=1 and sample size 500

Figure 3.8, shows that with 500 control and 500 treatment patients, the Type I error rate aligns with the anticipated significance level from the second look for censoring rates of 0.1 and 0.5. However, for a censoring rate of 0.9, the Type I error does not consistently meet the expected significance level until the sixth look.

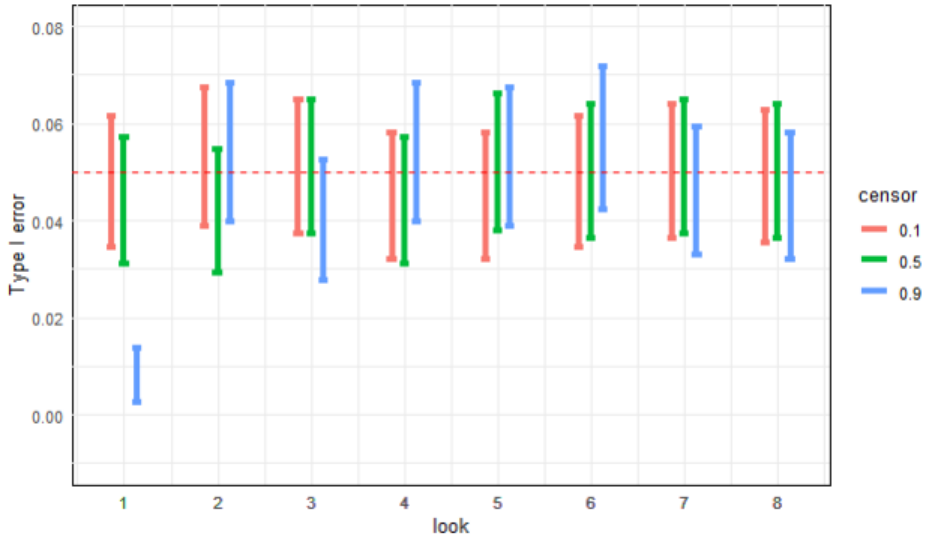


Figure 3.9: Type I error rate ± simulation error for HR=1 and sample size 1,000

In Figure 3.9, it is evident that with 1,000 control and 500 treatment patients, the Type I error rate aligns with the anticipated significance level from the first look for censoring rates of 0.1 and 0.5. However, for a censoring rate of 0.9, the Type I error consistently does not meet the expected significance level until the second look.

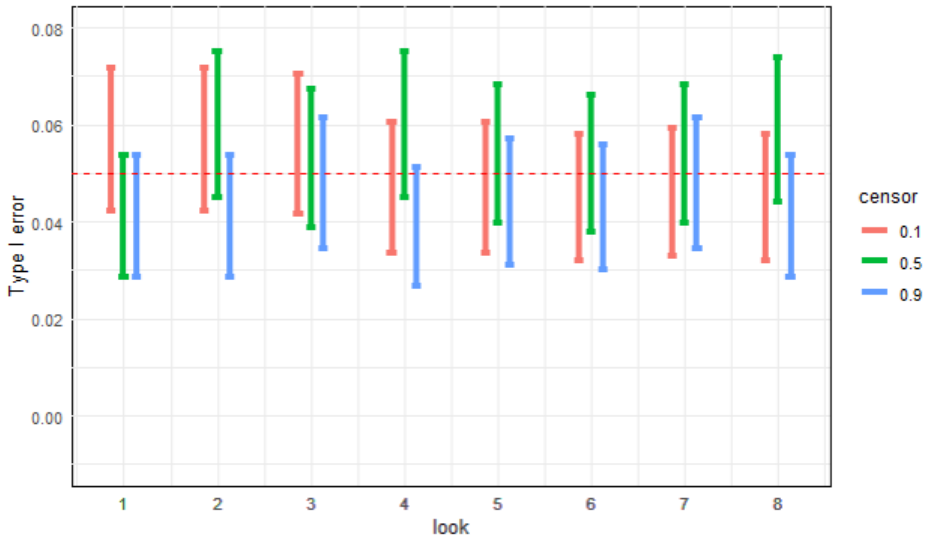


Figure 3.10: Type I error rate ± simulation error for HR=1 and sample size 5,000

Figure 3.10 shows that with 5,000 control and 5,000 treatment patients, the Type I error rate aligns with the anticipated significance level from the first look for all censoring rates from 0.1 to 0.9.

From the findings, it is evident that sample size plays a significant role in influencing the Type I error rate, with smaller sample sizes resulting in lower error rates than specified. Conversely, larger sample sizes show minimal susceptibility to Type I errors caused by censoring. For smaller sample sizes, it becomes imperative to account for censoring levels to maintain the desired Type I error rate, especially during the early interim analyses. Note that while we cannot directly control censoring, adjusting our statistical methods to account for it can help address potential biases.

Having obtained the Type I error at each look, the next step is to determine

the overall Type I error. This involves simulating data under a hazard ratio of 1, representing the null hypothesis, and examining the proportion of simulations with statistically significant outcomes for at least one look. This analysis offers a comprehensive understanding of the overall Type I error.

Generating data through simulation with a sample size of 5,000 patients for both treatment and control groups, alongside an overall censoring rate of 0.1, and conducting the study over a recruitment period and maximum duration of 365 days, yields an overall Type I error rate of 0.18. In conventional practice, multiple testing is typically conducted on independent samples. However, in this study, multiple tests are performed on cumulative data. Each dataset contains similar information but may vary in the number of patients and/or the duration of follow-up for some individuals. Consequently, due to the cumulative nature of the data, the anticipated overall Type I error rate differs significantly from what would be expected with independent samples.

Table 3.4: Report Parameters

| Number of Repeated Tests at 5% Level | Overall Type I Error for Independent Samples | Overall Type I Error for Accumulative Data |
|---|---|---|
| 1 | 0.05 | 0.05 |
| 2 | 0.09 | 0.08 |
| 3 | 0.14 | 0.11 |
| 4 | 0.19 | 0.13 |
| 5 | 0.23 | 0.14 |
| 10 | 0.40 | 0.19 |

Table 3.4 presents the expected overall Type I error rates for both independent samples and cumulative data at a 5% significance level. The overall type I error for independent samples, $\alpha^*$, was obtained by $\alpha^* = 1 - (1 - \alpha)^k$, where k is the number

of independent samples. The overall Type I error for cumulative data was empirically computed in the paper by Armitage et al. (1969) [11]. Notably, the value of 0.18 obtained from our simulations for 8 repeated tests falls between the expected values of 0.14 and 0.19 for 5 and 10 repeated tests on accumulative data, respectively.

# 4 FINAL REMARKS AND FUTURE WORK

In this thesis, we explored survival analysis, with a particular emphasis on interim statistical analyses within observational studies. The outcome involved the development of novel methodologies, including generating a simulation function and statistical analysis tools. Through these endeavors, we aimed to understand fundamental questions regarding the power and Type I error rates for interim analyses within observational studies.

The first major accomplishment was the creation of a function capable of simulating survival data following a Weibull distribution, and meeting the assumptions of the Cox proportional hazards model. The simulation function generated allows us to manipulate parameters such as the hazard ratio, overall censoring, maximum study duration, and sample size. This foundational tool laid the groundwork for subsequent analyses and investigations.

Additionally, we devised a function to perform statistical analyses at multiple interim points for any given dataset. This function requires the entire dataset under study as input, along with the desired number of looks to be performed. The function takes the input data and adjusts it to the data that would have been available at the specified time points. Upon execution, the function generates a summary table presenting the results obtained at each preliminary analysis. This tool enabled us to conduct in-depth examinations of the power and Type I error rates at various stages of observational studies, providing insights into the dynamics of these critical metrics.

One of the primary questions from our analysis was the determination of the power of tests conducted at each interim analysis, and whether input parameters exerted any influence on the power of preceding interim analyses. Through a meticulous

assembly of data and rigorous simulations, we were able to quantify known trends and relationships. Notably, we verified that power increases with sample size, the look number, and the value of the hazard ratio (as far away as the hazard ratio gets from 1), while decreasing with increased overall censoring. However, we also encountered an unexpected finding: after the conclusion of patient enrollment in the study, there is typically only a marginal increase in power.

Moreover, our investigation into the influence of sample size and overall censoring rate on the Type I error rate yielded insightful results. We found that smaller sample sizes tend to lead to more conservative tests, necessitating adjustments to censoring levels to maintain the desired Type I error rate. Note that while we cannot directly control censoring, adjusting our statistical methods to account for it can help address potential biases. Conversely, larger sample sizes exhibit minimal susceptibility to Type I errors caused by censoring.

Furthermore, our exploration extended to determining the overall Type I error by simulating data under a hazard ratio of 1. This analysis provided a comprehensive understanding of the cumulative Type I error across multiple interim analyses, thereby offering valuable insights into the overall integrity of the studies. The value of 0.18 for the overall Type I error obtained from our simulations with 8 repeated tests demonstrates that the multiple looks approach inflates the overall Type I error rate. To control this inflation, we should employ appropriate statistical methods such as adaptive designs. Looking forward, the next phase of this research should focus on attempting to control the overall Type I error using adaptive designs while maintaining the best statistical power. We aim to optimize the efficiency of observational studies while preserving the statistical validity of interim analyses.

In summary, this thesis has contributed significant advancements to the field of survival analysis. Through the creation of simulation functions, statistical analysis tools, and comprehensive reports, we have provided researchers with valuable re-

sources for designing and analyzing interim analyses within observational studies. Moving forward, we remain committed to furthering our understanding and improving methodologies in this critical area of research.

## BIBLIOGRAPHY


[1] *Illustration adapted from model displayed in "evidence-based practice in health".*, Retrieved from University of Canberra Library. (2023).

[2] *What is observational study desing and types*, Retrieved from Elsevier Library. (2023).

[3] Laura E et al. Bothwell, *Adaptive design clinical trials: a review of the literature and clinicaltrials.gov*, BMJ open **8,2** (2018).

[4] David Collett, *Modelling survival data in medical research, 3rd ed.*, Chapman Hall/CRC, 2014.

[5] Alexander Lorraine K. et al., *Causality*, Retrieved from UNC Gillings School of Global Public Health Library. (2015).

[6] Alexander Lorraine K et al., *Cohort studies*, Retrieved from UNC Gillings School of Global Public Health Library. (2015).

[7] Loprinzi CL et. al., *Prospective evaluation of prognostic variables from patient-completed questionnaires*, North Central Cancer Treatment Group. J Clin Oncol. (1994).

[8] Lise Frappier-Davignon Hill, N. and Brenda Morrison., *The periodic health examination.*, Can Med Assoc J 121 (1979), 1193–254.

[9] Rajiv Mahajan and Kapil Gupta, *The periodic health examination. canadian task force on the periodic health examination.*, Canadian Medical Association Journal **121,9** (1979), 1193–1254.

[10] Rajiv Mahajan and Kapil Gupta, *Adaptive design clinical trials: Methodology, challenges and prospect*, Indian journal of pharmacology **42,4** (2010), 201–207.

[11] C. K. McPherson P. Armitage and B. C. Rowe, *Repeated significance tests on accumulating data.*, Journal of the Royal Statistical Society **132** (1979), 235–244.

[12] Paul Rosenbaum and Donald Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrilka **70** (183), 41–55.

## VITA

Quentin Eloise was born on March 27th, 1997 in Lamentin, Martinique. He graduated from Lycée Bellevue in 2015. After high school, he attended Benedict College and received a Bachelor's degree in Applied Mathematics in 2019. In 2022, he attended Stephen F. Austin to pursue a Master of Science in Mathematics with an emphasis on statistics with plans to graduate in May 2024.

Permanent Address: 42 Rue Paulo Rosine Ravine Vilaine
Fort-de-France, 97200 Martinique

The style manual used in this thesis is A Manual For Authors of Mathematical Papers published by the American Mathematical Society.

This thesis was prepared by Quentin Eloise using LaTeX.