Stockholms universitet

Impact of exposure definition on exposure-outcome associations

Oliver Ng

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2023:7
Matematisk statistik
Juni 2023

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2023:7**
http://www.math.su.se

Impact of exposure definition on exposure-outcome associations

Oliver Ng*

June 2023

Abstract

Determining the association between a treatment's effect and time exposure is of great importance in regards to having a proper understanding of the effects, both harmful and beneficial, in the population. However, the definition of exposure time during treatment varies depending on how the user defines these periods. To answer the question on the impact of exposure definition, we make use of the Cox proportional hazards model to quantify the difference and accuracies of three common definitions, and in addition, for comparison with a known association, a simulated data set is used. The initial model with the real data set showed that the three definitions showed differing results, while the two Cox models returned equivalent results with emphasis on how the users handled certain parameters, such as the gap between refills. The simulated data further supports this and also shows that given extreme or non-ideal conditions, some of the definitions return wildly differing results or models with no statistical power.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: Obvious.spycrab@gmail.com. Supervisor: Daniel Ahlberg.

Abstract

Determining the association between a treatment's effect and time exposure is of great importance in regards to having a proper understanding of the effects, both harmful and beneficial, in the population. However, the definition of exposure time during treatment varies depending on how the user defines these periods. To answer the question on the impact of exposure definition, we make use of the Cox proportional hazards model to quantify the difference and accuracies of three common definitions, and in addition, for comparison with a known association, a simulated data set is used. The initial model with the real data set showed that the three definitions showed differing results, while the two Cox models returned equivalent results with emphasis on how the users handled certain parameters, such as the gap between refills. The simulated data further supports this and also shows that given extreme or non-ideal conditions, some of the definitions return wildly differing results or models with no statistical power.

Supervisor: Marie Linder, Daniel Ahlberg, Diego Hernan Giunta

Contents

Foreword

I'd like to thank Karolinska Institutet for letting me work on my Master's thesis on their behalf. Special thanks to my supervisors, Marie Linder and Diego Hernan Giunta, which provided me with vital contextual background for the thesis and Daniel Ahlberg from Stockholm University for agreeing to supervise and also lending some crucial suggestions.

1 Introduction

There's always been a need for statistical inferences in a pharmaceutical setting, and while the topic of the treatment's efficacy is probably at the forefront, another part of this is the epidemiology of its side effects, such as their proliferation or the association with the duration of the treatment.

For the data set, one can make use of the database held by the national board of health and Welfare which contains, among others, the prescription register present in most health care systems and likely contains information regarding the substance, dispensation date, and its amount.

However, even with those parameters mentioned, it could still lack some vital information central to a good statistical analysis, namely the explicit time the individual is using the drug, i.e,. the period of exposure to potential side effects. Thus, one needs to find a suitable way to estimate this period; an inability or poor estimation of this period can lead to the inclusion of serious bias in the ensuing results.

To address this issue, many ways to define this period have been suggested [6], ranging from the straightforward amount of defined daily dosage (DDD) times the amount dispensed[1] to the more statistically data driven method applying Waiting Time Distribution, a sort of stochastic process [3].

To highlight a specific method relevant to this project, there's a previously commonly used strategy designed to test the efficacy of a new treatment in Randomized Clinical Trials (RCT). The recommended method is to use what is called intention to treat (ITT) [2] which is biased towards the null effect, meaning it's a conservative method to observe the pharmacological exposure and its efficacy.

Another method of note is PP or Per protocol also called On Treatment, which is an exposure definition where instead one assigns the outcomes to a specific drug routine, which in consequence reduces the number of participants but also makes sure that the drug of interest is in focus. This is usually used when one wants to prove that a given drug is at least as potent as the comparator drug.

Despite the existence of numerous studies defining a standardized exposure duration, there is a lack of focus on comparing the numerous definitions against each other.

1.1 Main Thesis

The main suggestion for this thesis is to investigate how assigning the drug exposure period affects the association between exposure period and time to side effects. We compare the drug exposure period of three different definitions using both a real data set and a simulated data set with varying conditions.

1.2 Outline of the Thesis

This thesis consists of seven sections, with its own collection of subsections. We begin by elaborating on the background in **Chapter 2** to the main question and also showing how the data set is processed so that it's usable for our question. Then we give a short overview and the method to process the data set in **Chapter 3**. **Chapter 4** defines the numerous mathematical concepts used or that are useful for applying a Cox mode. In **Chapter 5** the models and their properties are elaborated, while the simulation part is reserved for **Chapter 6**. We then finally present the results in **Chapter 7**, concluding the thesis with a discussion in **Chapter 8** and making final remarks in **Chapter 9**.

2 Background

In this section, we'll be elaborating some of the background and contextualizing the central keypoints in the thesis.

2.1 Assigning the exposure duration

For this thesis, there's three ways to estimate the duration of drug exposure.

- Intention to treat (ITT)
- On-treatment (OT)
- Time-varying method (TV)

In ITT, we assume that the exposure episode starts at the dispensation date; in other words, as soon as the patient is dispensed the medication, it's considered in use, and it's only over either when an event such as a side effect is observed or the study ends disregarding any switches or halts in the treatment. This makes ITT one of the simpler ways to define this period.

OT otherwise known as On-Treatment or On Protocol, where instead of considering the drug in use as soon as possible, we emphasize its usage period. As such, exposure only starts when the treatment initiates and ceases when either the individual switches drugs, experiences an event, or the study ends. As a result, the number of observed events at the end is lower compared to ITT as we will see soon. Many patients don't stay on a single drug. In other words, we emphasize an uninterrupted drug regime for this definition.

The last definition is the TV method, where the individual contributes its exposure period to multiple groups. Say that the patient starts out on treatment A and then switches to treatment B. In that case, the individual contributes its exposure period to two different drugs; in this case, the exposure duration becomes a time variate due to it aggregating multiple parallel individuals at once. However, even if it contributes to multiple exposure periods, any events are associated with the individual's current drug regime, no matter the starting or any prior medication.

We return to these definitions when we apply them to the relevant data set.

2.2 Defined Daily Dosage

The defined daily dosage (DDD) is the maintenance dosage per day for a given drug; in the prescription data set, it refers to the expected amount of days the dispensation is assumed to last. So a DDD of 20 would mean to enough dosage for 20 days; the period outside of the expected DDD period where the individual hasn't refilled their prescription is referred to in this thesis as the gap period.

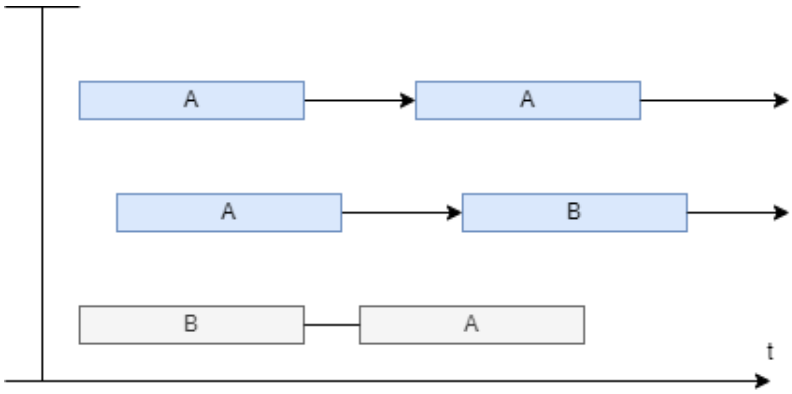

Figure 1: Visual representation of ITT method of two different treatment methods on three individuals, where the blue bars represent exposure period. Note how that blue bar persists even when the individual switches treatment at the end of the study.

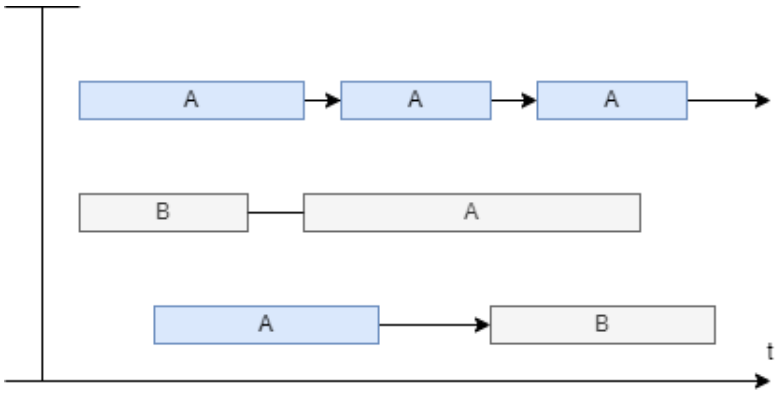

Figure 2: Diagram of OT method for three individuals, notice that when one switches treatment, the exposure period ceases and is thus considered censored.

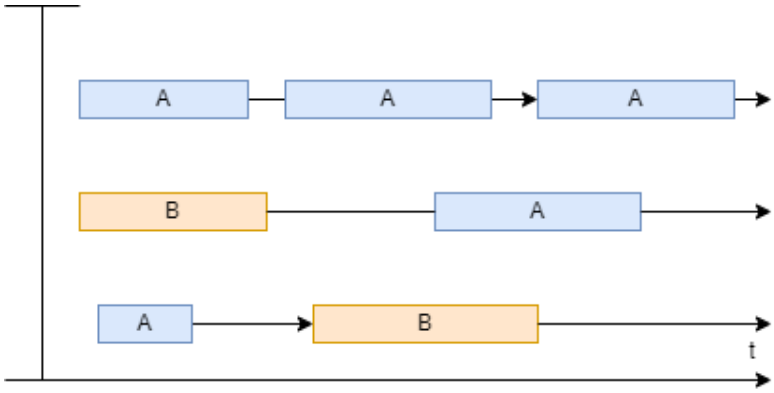

Figure 3: The TV method visualized for two individuals, where the blue bars represent exposure periods for A and the orange bars for B.

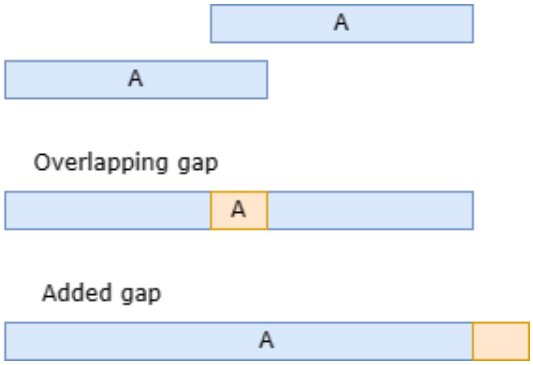

Figure 4: Example of how one can handle overlapping exposure periods. The blue bar is the expected exposure, while the orange bar represents the accounting change made by the overlapping gap.

2.3 Overlapping Dosage

Sometimes, the individual might have been dispensed another set of drugs during their expected dosage duration due to either losing track of their dispensed drug, wishing to stock up, or other unforeseen events. Thus, we are faced with the problem of handling overlapping dosages.

As suggested by the source[5] for anti-depressants, there are two ways to handle this issue. The first is to treat it as a singular exposure period, adding the overlap at the end. The other way is to treat it again as the same but end it at the expected dosage of the latest dispensing date. In the case of this paper, we reset the DDD period when a prescription dispensation is detected.

2.4 Medical Setting

The medical condition of interest is *plaque psoriasis* and its proposed treatment, *Rizankisumab* other close derivatives. The condition labeled as plaque psoriasis is a skin condition where parts of the skin harden into pink like inflammation, forming a plaque, hence its name. It's by no means severe, but it's noted to affect the individual's lifestyle for the foreseeable future.

The treatment in question, *Rizankisumab*, is an immunoglobin G1 monoclonal antibody. Its proposed usage is treating adult patients for psoriasis and psoriatic arthritis with moderate to severe active Crohn's disease that the patient has either lost or was intolerant to conventional therapy. The dose regime interval is a dose at weeks 0, 4, and 8, with a smaller dose at week 12 and every eight weeks thereafter. In addition, there are 47 other drugs in the data set, with some sharing the suffix "zumab" and being in differing medical types used to treat psoriasis. [9]

The outcome we'll be looking for is MACE (Major Adverse Cardiovascular Events), a collection term for any symptoms that display any abnormalities in the cardiovascular area, such as stroke, nonfatal myocardial infarction, and potentially cardiovascular death. Any prior MACE before the study doesn't count but is used as a covariate in the model.

2.5 Causation

The direct causation of any outcome, intentional or not, is most likely also affected by other underlying factors. In medical research, factors such as age, weight, and sex are underlying factors inherent in the individual that may have a negative or positive correlation that can be observed.

In addition, there are some less inherent factors not directly related to the individual, as prior treatment with differing drugs or exposure to other diseases could play such a role as well. Going even deeper, the correlation might even go to the genetic level to a degree we can't even describe or give a proper value to.

The above is meant to contextualize the fact that even with the most rigorous analyses of any covariate, there might still be some correlation not being considered.

3 Data Set

3.1 Drug and Patient Registers

The Swedish Prescribed Drug Register (PDR) held by Socialstyrelsen contains the data for all the prescribed drug dispensation in Sweden, its amount and the

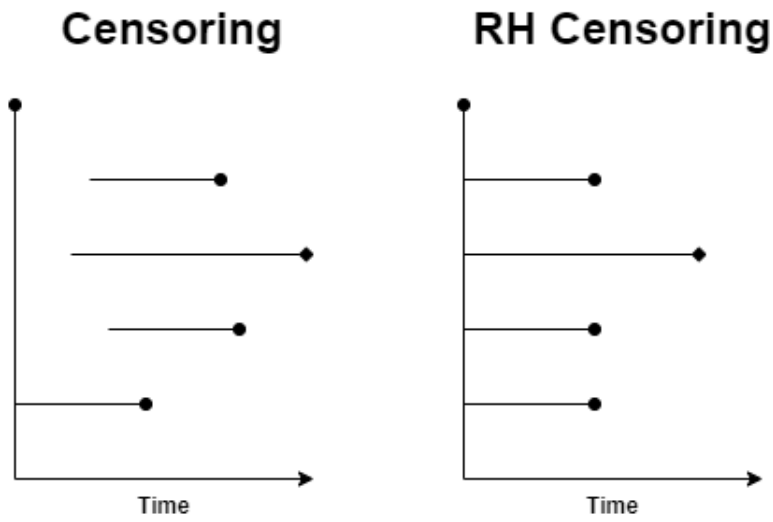

Figure 5: Diagram showing right-hand censoring, where we only concern ourselves with events after the study starts and how we can view those events as the amount of days from the start. The rounded points denote an event, and the square points are an individual that reached the end of the study.

recommended DDD but not any specific information regarding the usage of said drug. The data set we be using is a smaller set with the drugs used to treat psoariasis with some of the drugs sharing the -zumab suffix.

As a companion piece to the register is the collection of various outcomes from the National Patient Register (NPR) again sourced from Socialstyrelsen, it's a massive data set for all registered patients nation wide containing personal data such as security number, Age and gender, any admission to healthcare facilities, main diagnosis and other health conditions for each patients. [4]

Being the most comprehensive data set concerning all kinds of ailment, the NPR data set is pruned such that we only have the patient relevant to the PDR set. With some choice parameters which we elaborate later on **Chapter 5** .

Both data set start from 2005-07-05 and ends at 2022-12-31, due to the data set we use, we only concern ourselves with right hand censoring. The number of participants is 57764 and there were 3815 events registered.

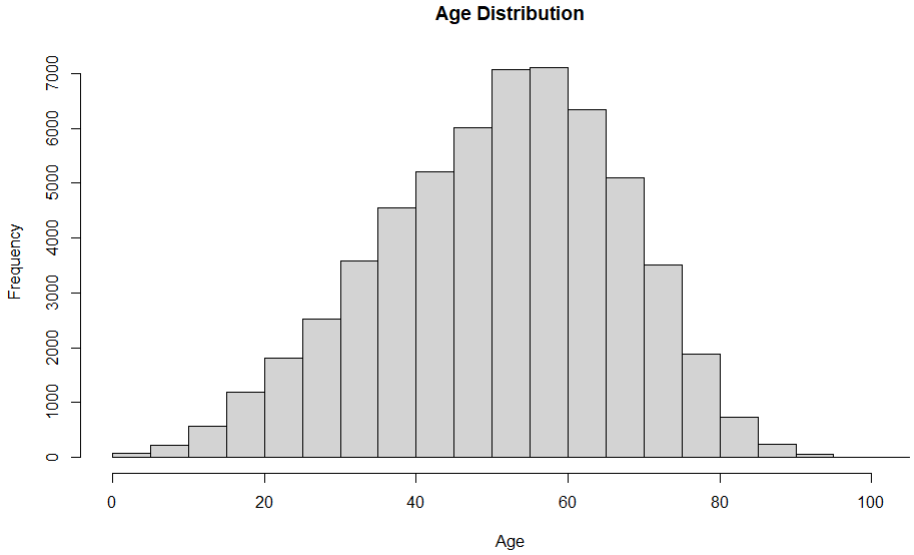

Figure 6: Distribution of Age in the data set, showing the general, older population present in the data set.

3.2 Classifying the drugs

The 48 drugs used for this study are additionally subdivided into smaller categories based on their medical effects as noted by Figure 3.2. The four main categories are *Other-IL*, *Non-Biologic*, *TNF-Inhibitors*, *Risankizumab* by itself, and *Non-exposed* for the remainder that's not part of the -zumab and could be considered a kind of control group. This batching makes the model more manageable further down the thesis. A clarification on the specifics of the drugs in question can be seen in Section 11.4 at page 58

3.3 Switching between treatments

An overview of the data set shows that not many stay on the same drug for too long, with some of the individuals rapidly switching between two treatments, as inferred from figure 9. While this is of no concern for ITT as it only concerns the initial drug, it could misrepresent the data if such switches occur often. It's central to the exposure assignment for OT where it's censored after any switches, and TV method where any prolonged period outside of the expected DDD period is considered a **Non-exposed** period.

Unless stated otherwise, for any treatment that has a refill period longer than the expected DDD duration, we consider any day over 15 as **Non-exposed**

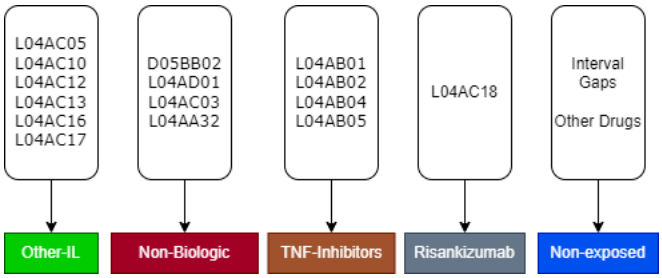

Figure 7: Diagram for the categorization of each individuals drugs.

period.

3.4 Processing the data set

Due to the varying degree of complexity, each definition requires some specific data processing methodology. In the context of survival analysis, it's needed for us to extract the delay between start to end, the outcome, or the censoring of the particular individual, which makes up most of the complication for each definition.

As we noted in the background section, there are some core differences between the definitions that we'll now apply to the relevant data set.

3.4.1 Core Concept

For us to process a Cox-friendly data set, we use the same main idea for all three, with some noted differences for each one.

In general, we aggregate the multiple DDD periods as a single uninterrupted drug period, including any gap periods that are under 15 days; for *ITT* this isn't as relevant compared to its counterparts.

3.4.2 Intention To Treat

This is the most straightforward of the definitions, as we only care about whether the individual began its treatment with a specified drug. In such a case, we ignore all other treatments and also any changes, swaps, or if the individual leaves the group entirely.

By data processing, we simply take the total delay between the start and the outcome or end of the study for each participant, disregarding any complicated assumptions.

3.4.3 On Treatment

On-Treatment requires some more finesse, as we want an uninterrupted treatment. As per the definition stated earlier, if any switch between medications is detected, the individual is considered censored regardless of their true outcome.

Of equal importance for the OT definition are the gaps between each refill of the drug, which, with a long enough gap in the exposure period, could be seen as no longer following the drug treatment. In this case, and for the TV method as well, any extended gap period is considered a different drug exposure period and censored.

By data processing, we start out the same as ITT until we notice any interruptions. This means that when we detect that either the individual is taking a new drug or there's an extended gap period, we assume that the treatment is interrupted and thus censored henceforth.

3.4.4 Time Varying

Since we let each individual contribute "as treated", we now have to setup the model with exposure being a time-varying variable; in other words, we divide the individual into sections for each treatment, making it a time-dependent model. In such a case, any given individual could contribute to its exposure period for all potential exposures where the event is associated with its current drug regime or if it is discovered in the gap period. Though note that since any prior drug has no bearing on the current medication by this definition, this could be a source of miss-classifications.

Mechanically, we don't censor the individual for each new drug regime or extend the gap between the prescription dates. As an example, if we have drug 1 at time 0 and then switch to drug 2 at time t_i, we assign it as an interval of $[0, t_i]$ for drug 1 and then check the next interval $[t_i, t_j]$ for drug 2 until another switch or an extended gap-period is detected, where it's assigned as **Non-exposed**.

This turns the data set into an interval-based data set that we can use for a time-dependent Cox model.

3.5 Coefficients

As a semi-parametric model, the Cox model requires some coefficients, which are also of interest in regards to how each definition interacts with the coefficients.

However, since we're more interested in the effect of the hazard ratio between the definitions than the actual efficacy of the treatment, we'll include a smaller selection of covariates than one would expect, with the ones included used to compare the hazard ratio in the coefficients.

The covariates of interest are **Age**, **Sex**, specific prior health conditions, and prior drug usage as suggested by the background section, such as prior **MACE** symptoms, **Hypertension** and **Obesity**.

4 Mathematical Theory

In this section, we properly introduce the mathematical concept used in this thesis. Note that unless specifically stated, the description of the concepts is based on *Survival and Event History Analysis* [10] using the same notation.

4.1 Supporting Concepts

Before we elaborate on the main tool used for this thesis, we need to state some supporting concepts that are vital to the Cox proportional hazard model.

4.1.1 Survival function

Consider a data set of an n-sized population. We want to estimate the occurrence or rate of an event of interest for the individuals in the population; this is referred to as the hazard rate. In this case, we denote $\alpha(t)$ to specify the hazard rate of an instanteous probability that an event in the population will occur in the time frame $[t, t + dt)$ given that it has not occurred earlier:

Without assuming any parametric assumption, the hazard rate $\alpha(t)$ can be any nonnegative with the estimated cumulative hazard of $A(t) = \int_0^t \alpha(s)ds$ without any structure assumption. This is akin to estimating the cumulative distribution function, which leads to the Nelson Aalen estimator:

$$\hat{A}(t) = \sum_{T_j \leq t} \frac{1}{Y(T_j)} \tag{1}$$

Where $Y(T_j)$ is the amount of individual at risk before time T_j, to contextualize this results, if we say have 1.5 at time t that means that any individual at time t is 1.5 times more likely to experience given that no event has occurred prior to t.

Following the same data set, if we want to instead find out the probability that for a randomly selected individual the event will occur *after* time t we define the survival rate, we start by dividing the desired time interval $[0, t]$ into smaller K units of $[0 = t_0 < ... < t_K = t]$. Using the multiplicative rule of conditional probability we have

$$S(t) = \prod_{k=1}^{K} S(T_k|t_{k-1})$$

Such that each interval contains one event. This conditional probability can be estimated as $S(T_k|t_{k-1})$ as $1 - 1/Y(t_{k-1}) = 1 - 1/Y(T_j)$ and setting those into the estimate we obtain

$$\hat{S}(t) = \prod_{T_j \leq t} \left\{ 1 - \frac{1}{Y(T_j)} \right\} \tag{2}$$

Which is the Kaplan-Meier estimate. In contrast, rather than estimating the risk of a given individual at t, it instead estimates the probability that the individual hasn't experienced an event at t; in other words, if the survival function at t is 0.4, that states that 40% of the subjects survive to t.

Both of these estimates take censoring into account, as $Y(T_j) > 0$ implies that some individuals haven't experienced the event either by leaving the population or by exceeding the duration of the exposure period.

4.1.2 Martingale process

Let $M = \{M_0, , M_1, ...\}$ be a stochastic process in discrete time. The process M is called martingale if:

$$E(M_n|M_0, M_1, ..., M_{n-1}) = M_{n-1} \tag{3}$$

For each $n \geq 1$. In other words, the process is a martingale if the conditional expectation given in the past equals the previous value.

As a consequence, if given a sequence where we know that $E[M_0] = 0$, then we can show that:

$$E[M_n] = E[E(M_{n-1}|f_0)] = E[E(M_{n-1})] = E[M_0] = 0 \tag{4}$$

Where f_0 is the known history of outcomes prior to M_n. Applying this to a counting process $N(t)$

$$E\left\{M(t) - M(s)|f_s\right\} = E\left\{N(t) - N(s)\right\} \ - \lambda(t - s) = 0 \tag{5}$$

For $t > s$, we denote:

$$E\{M(t)|f_s\} = M(s)$$

Which illustrates that the process is indeed a martingale.

4.2 Cox Proportional Hazard Model

Consider a data set with n individuals where we count each event the population experiences. We want to model the correlation of each event to a number of covariates of interest using a regression model.

For the regression model, we have the counting processes $N_1, N_2, ..., N_n$ where $N_i(t)$ is the number of occurrences for individual i in $[0, t]$, with $N_i(t) =$

0 *or* 1 given the context of our survival analysis (events or no events). As we have a counting process, we have an accompanying intensity process defined as:

$$\lambda_i(t) = Y_i(t)\alpha(t|\boldsymbol{x}_i) \tag{6}$$

$Y_i(t)$ is the indicator variable if individual i is at risk for the relevant event, and $\alpha(t|\boldsymbol{x}_i)$ is the hazard rate defined conditional on the specific vector of covariates $\boldsymbol{x}_i$ for the individual i.

The Cox Proportional Hazard Model, called Cox model henceforth, is a common semi-parametric regression model in survival analysis. As such, by the nature of regression models, we assume that there's a $\boldsymbol{x}$ for a given individual i in relation to the hazard rate $\alpha(t|\boldsymbol{x}_i)$:

$$\alpha(t|\boldsymbol{x}_i) = \alpha_0(t)r(\boldsymbol{\beta}, \boldsymbol{x}_i(t)) \tag{7}$$

The term $r(\boldsymbol{\beta}, \boldsymbol{x}_i(t))$ is called the relative risk function, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is the regression coefficient describing the covariates in the model, and $\alpha_0(t)$ is the baseline hazard rate that is inherent in that model. If the relative term $r(\boldsymbol{\beta}, \boldsymbol{x}_i(t)) = 1$ then the baseline hazard rate corresponds to the individual hazard when the covariates are set to zero. Hence, it's called a semiparametric model, as we both have a nonparametric and a parametric part in the regression.

The relative risk function can be stated as an exponential variant, yielding:

$$r(\boldsymbol{\beta}, \boldsymbol{x}_i(t)) = exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t))$$

Thus, the risk regression model is now presented more in line with the typical regression model used for the Cox regression model, which has the added property of proportionality. Consider a Cox model with only one covariate, then consider what happens if we increase the effect of covariate x:

$$\begin{aligned} \alpha(t|x+1) &= \alpha_0(t)exp(\beta_1(x+1)) \\ &= \alpha_0(t)exp(\beta_1 x + \beta_1) \\ &= (\alpha_0(t)exp(\beta_1 x))\, exp(\beta_1) \\ &= \alpha(t|x)exp(\beta_1) \end{aligned} \tag{8}$$

We can see here that by increasing x by one or any other magnitude, the increase is proportional to the covariate. Stated differently:

$$\frac{\alpha(t|x+1)}{\alpha(t|x)} = exp(\beta_1) \tag{9}$$

We can see that it increases proportionally by β_1. Hence the inclusion of proportionality in its name, which well demonstrates the applications of this model. As an example between two subjects:

$$\frac{\alpha(t|x_i)}{\alpha(t|x_j)} = exp((x_i - x_j)\beta) \tag{10}$$

As such, the hazard ratio is proportional between the subjects if the other covariate is kept the same.

4.2.1 Estimation and Partial Likelihood

Due to the semi-parametric nature, directly using the likelihood method to estimate the parameters is not recommended. Thus, one needs to apply a different kind of Likelihood estimation.

We note that by combining the intensity process (6) and the hazard rate (7), we may write:

$$\lambda_i(t) = Y_i(t)\alpha_0(t)r(\boldsymbol{\beta}, \boldsymbol{x}_i(t)) \tag{11}$$

We introduce the aggregated counting process $N.(t) = \sum N_i(t)$ that registers all the events in the population. Note that this leads to the aggregated intensity processes $\lambda.$:

$$\lambda.(t) = \sum_i^n \lambda_i(t) = \sum_i^n Y_i(t)\alpha_0(t)r(\boldsymbol{\beta}, \boldsymbol{x}_i(t)) \tag{12}$$

This intensity process can be factorized into:

$$\pi(i|t) = \frac{\lambda_i(t)}{\lambda.(t)} = \frac{Y_i(t)\alpha_0(t)r(\boldsymbol{\beta}, \boldsymbol{x}_i(t))}{\sum_l^n Y_l(t)\alpha_0(t)r(\boldsymbol{\beta}, \boldsymbol{x}_l(t))} \tag{13}$$

So now we only need to find $\boldsymbol{\beta}$ such that it maximizes the partial likelihood:

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{T_j} \pi(i|t) = \prod_{T_j} \frac{r(\boldsymbol{\beta}, \boldsymbol{x}_i(T_j))}{\sum_{l \in R_j} r(\boldsymbol{\beta}, \boldsymbol{x}_l(T_j))} \tag{14}$$

Where $R_j = \{l|Y_l(T_j) = 1\}$ is the *risk set* of individuals at risk at T_j where $t < T_j$. The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is the value of which $\boldsymbol{\beta}$ maximizes (14); in Section 4.2.3, we show that this estimator enjoys large sample properties similar to ordinary maximum likelihood estimators.

4.2.2 Cumulative hazard and survival probabilities

Mentioned earlier, we had noted that α_0 is the baseline hazard, this can also be estimated as the cumulative hazard rate of $A_0(t) = \int \alpha_0(u)du$.

$$\lambda.(t) = \left(\sum_{i=1}^n Y_i(t)r(\boldsymbol{\beta}, \boldsymbol{x}_i(t)) \right) \alpha_0(t) \tag{15}$$

Chances are that we don't know the true $\boldsymbol{\beta}$, we'd have to estimate the Nelson-Aalen estimation by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ and estimating the hazard rate over the aggregated count:

$$\hat{A}_0(t) = \int_0^t \frac{dN.(u)}{\sum_{i=1}^n Y_i(u) r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i(u))} = \sum_{T_j \leq t} \frac{1}{\sum_{l \in R_j} Y_i(u) r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i(T_j))} \tag{16}$$

Where $N. = \sum_{i=0}^{\infty} N_i(t)$ is the aggregated counting process. This gives us an estimator for the cumulative baseline hazard; this is also called the Breslow estimator. When the covariates are fixed, we have an adjusted estimation corresponding to the covariate vector x_0:

$$A(t|\boldsymbol{x}_0) = \int_0^t \alpha(u|\boldsymbol{x}_0) du = r(\boldsymbol{\beta}, \boldsymbol{x}_0) A_0(t) \tag{17}$$

This is then estimated similarly as:

$$\hat{A}(t|\boldsymbol{x}_0) = r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_0) \hat{A}_0(t) \tag{18}$$

Conversely, if some of the covariates are time-dependent, it's not as meaningful to estimate the cumulative hazard for a fixed value of a covariate vector. As such, one can instead estimate the cumulative hazard in an interval of $[0, t]$ given a corresponding path $\boldsymbol{x}_0 : 0 < s \leq t$. The path to such a cumulative hazard corresponds to:

$$A(t|\boldsymbol{x}_0) = \int_0^t r(\boldsymbol{\beta}, \boldsymbol{x}_0(u)) \alpha_0(u) du \tag{19}$$

And again, this can be estimated as:

$$\hat{A}(t|\boldsymbol{x}_0) = \int_0^t r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_0(u)) d\hat{A}_0(u) = \sum_{T_j \leq t} \frac{r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_0(T_j))}{\sum_{l \in R_j} r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_l(T_j))} \tag{20}$$

The survival function can be estimated in a similar way to the hazard rate corresponding to either non- or time-based covariates. As such, we can write the aforementioned estimation as the corresponding product integral denoted Π:

$$S(t|\boldsymbol{x}_0) = \Pi \left\{ 1 - dA(u|\boldsymbol{x}_0) \right\} \tag{21}$$

Which then can be estimated with $\boldsymbol{\beta}$ for the estimator

$$\hat{S}(t|\boldsymbol{x}_0) = \Pi \left\{ 1 - d\hat{A}(u|\boldsymbol{x}_0) \right\} = \prod_{T_j \leq t} \left\{ 1 - \Delta \hat{A}(T_j|\boldsymbol{x}_0) \right\} \tag{22}$$

Which is an estimate that's asymptotically normally distributed around its true value, with variance estimated in Section 4.2.4.

4.2.3 Large Sample properties of $\hat{\beta}$

The following two sections go deeper into the asymptomatic distribution, justifying the variance estimation. The following sections, 4.2.3 and 4.2.4, are included for posterity's sake and can be skipped.

In this subsection, we outline the derivation used to prove that the above mentioned equation for the cox model and $\hat{\boldsymbol{\beta}}$ is indeed multivariate normally distributed around the true $\boldsymbol{\beta}$ when it's used in larger data set. Its covariance matrix is also in extension the expected information matrix.

Keep in mind that prior to this section, $\boldsymbol{\beta}$ was denoted for the true value of the covariates as well as for the partial likelihood estimates; however, when handling large sample sizes, it is to one's advantage to distinguish the two uses, so we define $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta}$ as the true and partial likelihood, respectively.

As an opening, we define the log likelihood of for the upper time limit τ:

$$l_{Cox}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \boldsymbol{\beta}^T \boldsymbol{x}_i(u) - log \boldsymbol{S}^0_{Cox}(\boldsymbol{\beta}, u) \right\} dN_i(u) \tag{23}$$

Where we define the following notations:

$$\begin{aligned} \boldsymbol{S}^0_{Cox}(\boldsymbol{\beta}, t) &= \sum_{l=1}^{n} Y_l(t) exp\left\{ \boldsymbol{\beta}^T \boldsymbol{x}_l(t) \right\} \\ \boldsymbol{S}^1_{Cox}(\boldsymbol{\beta}, t) &= \sum_{l=1}^{n} Y_l(t) \boldsymbol{x}_l(t) exp\left\{ \boldsymbol{\beta}^T \boldsymbol{x}_l(t) \right\} \\ \boldsymbol{S}^2_{Cox}(\boldsymbol{\beta}, t) &= \sum_{l=1}^{n} Y_l(t) \boldsymbol{x}_l^{(+)}(t) exp\left\{ \boldsymbol{\beta}^T \boldsymbol{x}_l(t) \right\} \end{aligned} \tag{24}$$

Where $(+)$ denotes where the specific matrix vv^t given a column vector $v^{(+)}$. Score function then follows:

$$\boldsymbol{U}_{Cox}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}^T} l_{Cox}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \boldsymbol{x}_i(u) - \frac{\boldsymbol{S}^1_{Cox}(\boldsymbol{\beta}, u)}{\boldsymbol{S}^0_{Cox}(\boldsymbol{\beta}, u)} \right\} dN_i(u) \tag{25}$$

With its observed information matrix defined as:

$$\boldsymbol{I}_{Cox}(\boldsymbol{\beta}) = -\frac{\partial}{\partial \boldsymbol{\beta}^T} \boldsymbol{U}_{Cox}(\boldsymbol{\beta}) = \int_0^{\tau} \frac{\boldsymbol{S}^2_{Cox}(\boldsymbol{\beta}, u)}{\boldsymbol{S}^0_{Cox}(\boldsymbol{\beta}, u)} - \left(\frac{\boldsymbol{S}^1_{Cox}(\boldsymbol{\beta}, u)}{\boldsymbol{S}^0_{Cox}(\boldsymbol{\beta}, u)} \right)^{(+)2} dN.(u) \tag{26}$$

Following this, we denote a new notation such that:

$$\boldsymbol{V}_{Cox}(\boldsymbol{\beta}, t) = \frac{\boldsymbol{S}^2_{Cox}(\boldsymbol{\beta}, u)}{\boldsymbol{S}^0_{Cox}(\boldsymbol{\beta}, u)} - \left(\frac{\boldsymbol{S}^1_{Cox}(\boldsymbol{\beta}, u)}{\boldsymbol{S}^0_{Cox}(\boldsymbol{\beta}, u)} \right)^{(+)2}$$

If we now insert $\boldsymbol{\beta_0}$ in (**25**) and use the decomposition:

$$dN_i(u) = \lambda_i(u)du + dM_i(u) = Y_i(u)\alpha_0 exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(u))du + d_i(u)$$

Where M_i represents the martingale process. With some algebra, we find that:

$$\boldsymbol{U}_{Cox}(\boldsymbol{\beta}_0) = \sum_{i=1}^{n} \int_0^\tau \left\{ \boldsymbol{x}_i(u) - \frac{\boldsymbol{S}^1_{Cox}(\boldsymbol{\beta}_0, u)}{\boldsymbol{S}^0_{Cox}(\boldsymbol{\beta}_0, u)} \right\} dM_i(u) \tag{27}$$

The integrands presented are predictable processes, a process we know at a prior time, then the score function is a stochastic integral evaluated at the true $\boldsymbol{\beta_0}$. Particularly $E[\boldsymbol{U}_{Cox}(\boldsymbol{\beta_0})] = 0$. If we set the limit on right hand side at t we have a stochastic process. This process is a martingale with predictable variational process that, when evaluated at τ become:

$$< \boldsymbol{U}_{Cox}(\boldsymbol{\beta_0}) > (\tau) = \int_0^\tau \boldsymbol{V}(\boldsymbol{\beta_0}, u) S^{(0)}_{Cox}(\boldsymbol{\beta_0}, u)\alpha_0(u)du \tag{28}$$

The denotation $< M >$ is a predictable variation process (cf. Appendix 11.2). Note that the counting process $N.(t)$ has an intensity process $S^{(0)}_{Cox}(\boldsymbol{\beta_0}, t)\alpha_0(t)$, the observed information matrix at $\boldsymbol{\beta_0}$ becomes:

$$\boldsymbol{I}_{cox}(\boldsymbol{\beta_0}) =< U_{Cox}(\boldsymbol{\beta_0}) > (\tau) + \int_0^t V_{Cox}(\boldsymbol{\beta_0}, u)dM.(u) \tag{29}$$

Where $M. = \sum_{l=1}^{n} M_i$ is also a martingale. Following this, we can now show by Martingale central limit theorem (Appendix 11.3) that under suitable regularity conditions converges $n^{-1/2}\boldsymbol{U}_{Cox}(\boldsymbol{\beta_0})$ in distribution to a multivariate normal distribution with mean zero and covariance matrix Σ_{Cox}

By these results, we can show that the estimated $\hat{\boldsymbol{\beta}}$ follows in the same way, by Taylor expanding the score function:

$$0 = U_{Cox}(\boldsymbol{\beta}) \approx U_{Cox}(\boldsymbol{\beta_0}) - I_{Cox}(\boldsymbol{\beta_0})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \tag{30}$$

Then we can obtain:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \approx (n^{-1} I_{Cox}(\boldsymbol{\beta_0}))^{-1} n^{-1/2} U_{Cox}(\boldsymbol{\beta_0}) \approx \Sigma^{-1}_{Cox} n^{-1/2} U_{Cox}(\boldsymbol{\beta_0}) \tag{31}$$

It follows that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})$ converges in distribution to a multivariate normal distribution with mean zero and covariance matrix Σ^{-1}_{Cox}. Justifiably $\hat{\beta}$ then is approximately multivariate normal around β_0 with estimate covariance of the expected information matrix $I_{Cox}(\hat{\beta})^{-1}$

Following this argument to Cox model, although there's some minor modification for the relative risk function $r(\boldsymbol{\beta}, \boldsymbol{x}_i(u))$ but the formula as a whole becomes more complicated. As the score function becomes:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \frac{\dot{\boldsymbol{r}}(\boldsymbol{\beta}, \boldsymbol{x}_i)}{r(\boldsymbol{\beta}, \boldsymbol{x}_i(u))} - \frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, u)}{\boldsymbol{S}^{(0)}(\boldsymbol{\beta}, u)} \right\} dN_i(u) \tag{32}$$

Where $\dot{\boldsymbol{r}}(\boldsymbol{\beta}, \boldsymbol{x}_i) = \dfrac{\partial}{\partial \beta} r(\boldsymbol{\beta}, \boldsymbol{x}_i(t))$, we also define the following notations.

$$\begin{aligned} S^{(0)}(\boldsymbol{\beta}, t) &= \sum_{i=1}^{n} Y_l(t) r(\boldsymbol{\beta}, x_l(t)) \\ S^{(1)}(\boldsymbol{\beta}, t) &= \sum_{i=1}^{n} Y_l(t) \dot{\boldsymbol{r}}(\boldsymbol{\beta}, x_l(t)) \\ S^{(2)}(\boldsymbol{\beta}, t) &= \sum_{i=1}^{n} Y_l(t) \frac{\dot{\boldsymbol{r}}(\boldsymbol{\beta}, x_l(t))^{(+)2}}{r(\boldsymbol{\beta}, x_l(t))} \end{aligned} \tag{33}$$

$$\boldsymbol{V}(\boldsymbol{\beta}, t) = \frac{\boldsymbol{S}^{(2)}(\boldsymbol{\beta}, u)}{\boldsymbol{S}^{(0)}(\boldsymbol{\beta}, u)} - \left(\frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, u)}{\boldsymbol{S}^{(0)}(\boldsymbol{\beta}, u)} \right)^{(+)2} \tag{34}$$

We have the predictable variation process of the score function at β_0:

$$< \boldsymbol{U}(\boldsymbol{\beta_0}) > (\tau) = \int_0^{\tau} \boldsymbol{V}(\boldsymbol{\beta_0}, u) S^{(0)}(\boldsymbol{\beta_0}, u) \alpha_0 du \tag{35}$$

Finally, the observed information matrix $I(\boldsymbol{\beta})$ can be written as the sum of predictable variation process of the score and the stochastic integral. Thus, the expected information matrix is

$$\mathcal{I}(\boldsymbol{\beta}) = \int_0^{\tau} \boldsymbol{V}(\boldsymbol{\beta}, u) dN_{\cdot}(u) = \sum_{T_j} \boldsymbol{V}(\boldsymbol{\beta}, T_j) \tag{36}$$

There are some noted differences between the expected and observed information matrixes, but those quantities are for the most part negligible, though the expected information tends to be the more stable of the two.

4.2.4 Large Sample properties of Hazard and survival functions

In this section we again show that the hazard and survival function presented earlier is also normally distributed around the true value. Indeed we show this by first noting the difference:

$$\begin{aligned} \hat{A}(t|\boldsymbol{x}_0) - A(t|\boldsymbol{x}_0) = &\int_0^t r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_0(u)) (d\hat{A}_0(u; \hat{\boldsymbol{\beta}}) - d\hat{A}_0(u; \boldsymbol{\beta}_0)) \\ &+ \int_0^t r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_0(u)) (d\hat{A}_0(u; \boldsymbol{\beta}_0) - \alpha_0(u) du) \\ &+ \int_0^t \Big(r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_0(u)) - r(\boldsymbol{\beta}_0, \boldsymbol{x}_0(u)) \Big) \alpha_0(u) du \end{aligned} \tag{37}$$

And:

$$\hat{A}_0(t;\boldsymbol{\beta}) = \int_0^t \frac{dN.(u)}{S^{(0)}(\boldsymbol{\beta},u)} \tag{38}$$

Asymptotically we are able to replace the estimated relative risk function with $r(\boldsymbol{\beta}, x_0(u))$ for the first two terms on the right hand side. Then the first term by Taylor expansion we can approximate it:

$$-\int_0^t r(\boldsymbol{\beta_0}, x_0(u)) \frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta_0},u)}{\boldsymbol{S}^{(0)}(\boldsymbol{\beta_0},u)} dN.(u)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \tag{39}$$

On the second term, we use the Doob Meyer to decomposition (cf. Appendix 11.4) such:

$$dN.(u) = S^{(0)}(\boldsymbol{\beta_0},u)\alpha_0 du + dM.(u) \tag{40}$$

For the approximation

$$\int_0^t \frac{r(\boldsymbol{\beta_0}, x_0(u))}{\boldsymbol{S}^{(0)}(\boldsymbol{\beta_0},u)} dM.(u) \tag{41}$$

Finally the last term again by Taylor expansion approximates to

$$\int_0^t r(\boldsymbol{\beta_0}, x_0(u))\alpha_0 du(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \tag{42}$$

Then we have the final approximation of (37)

$$\begin{aligned} \hat{A}(t|\boldsymbol{x}_0) - A(t|\boldsymbol{x}_0) \approx & \int_0^t \frac{\dot{\boldsymbol{r}}(\boldsymbol{\beta}, \boldsymbol{x}_i)}{\boldsymbol{S}^{(0)}(\hat{\boldsymbol{\beta}},u)} dM.(u) \\ & + \int_0^t \left[\dot{\boldsymbol{r}}(\boldsymbol{\beta_0}, \boldsymbol{x}_i) - r(\boldsymbol{\beta_0}, x_0(u)) \frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta_0},u)}{\boldsymbol{S}^{(0)}(\boldsymbol{\beta_0},u)} \right] \alpha_0 du(\hat{\boldsymbol{\beta_0}} - \boldsymbol{\beta_0}) \end{aligned} \tag{43}$$

We can show that the first term is approximately normal distributed with mean zero and variance estimated by

$$\hat{\omega}^2(t|\boldsymbol{x}_0) = \int_0^t \left(\frac{r(\hat{\boldsymbol{\beta}}, x_0(u))}{\boldsymbol{S}^{(0)}(\hat{\boldsymbol{\beta}},u)} \right) dN.(u) = \sum_{T_j \le t} \left(\frac{r(\hat{\boldsymbol{\beta}}, x_0(T_j))}{\boldsymbol{S}^{(0)}(\hat{\boldsymbol{\beta}},T_j)} \right)^2 \tag{44}$$

This can be expanded further by defining the second term with $\hat{G}(t|\boldsymbol{x}_0)^T \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1} \hat{G}(t|\boldsymbol{x}_0)$ which $\hat{G}(t|\boldsymbol{x}_0)$ defined as:

$$\hat{G}(t|\boldsymbol{x}_0) = \int_0^t \left\{ \dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, x_0(u)) - r(\hat{\boldsymbol{\beta}}, x_0(1)) \frac{\boldsymbol{S}^{(0)}(\hat{\boldsymbol{\beta}},u)}{\boldsymbol{S}^{(0)}(\hat{\boldsymbol{\beta}},u)} \right\} d\hat{A}_0(u)$$

$$= \frac{\dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_o(T_j))}{\sum_{l \in R_j} r(\hat{\boldsymbol{\beta}}, x_l(T_j))} - \sum_{T_j \leq t} r(\hat{\boldsymbol{\beta}}, x_0(T_j)) \frac{\dot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_0(T_j))}{\left(\sum_{T_j \leq t} r(\hat{\boldsymbol{\beta}}, x_l(T_j))\right)^2}$$

It can be shown that the last two terms in (43) are asymptotically independent; it follows from that the $\hat{A}(t|\boldsymbol{x}_0)$ is approximately small around its true value with a variance that's estimated by

$$\hat{\sigma}^2(t|\boldsymbol{x}_0) = \hat{\omega}^2(t|\boldsymbol{x}_0) + \hat{G}(t|\boldsymbol{x}_0)^T \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1} \hat{G}(t|\boldsymbol{x}_0) \tag{45}$$

We lastly consider the converse for the survival function, given that it's quite similar to the hazard function we can see from:

$$\hat{S}(t|\boldsymbol{x}_0) - S(t|\boldsymbol{x}_0) =\approx -S(t|\boldsymbol{x}_0)(\hat{A}((t|\boldsymbol{x}_0)) - A(t|\boldsymbol{x}_0)) \tag{46}$$

It follows that the survival function $\hat{S}(t|\boldsymbol{x}_0)$ is also normally distributed around $S(t|\boldsymbol{x}_0)$ with a variance of $S(t|\boldsymbol{x}_0)^2$ times the variance of $\hat{A}(t|\boldsymbol{x}_0)$. Then the estimator for the variance of the survival function is:

$$\hat{\tau}^2(t|\boldsymbol{x}_0) = \hat{S}(t|\boldsymbol{x}_0)\hat{\sigma}^2(t|\boldsymbol{x}_0) \tag{47}$$

4.2.5 Martingale residual

Martingale residual is an approach to detecting or assessing nonlinearity in any continuous covariates. First, we introduce the cumulative intensity processes:

$$\Lambda_i(t) = \int_0^t \lambda_i(u)du = \int_0^t Y_i(u) r(\boldsymbol{\beta}, \boldsymbol{x}_i(u)) \alpha_0(u) du \tag{48}$$

Setting it with the estimated partial likelihood $\hat{\boldsymbol{\beta}}$ and the *Breslow estimate* we attain:

$$\hat{\Lambda}_i(t) = \int_0^t Y_i(u) r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i(u)) d\hat{A}_0(u) = \sum_{T_j \leq t} \frac{Y_i(T_j) r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_i(T_j))}{\sum_{l \in R_j} r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_l(T_j))} \tag{49}$$

And consequently, we have the *martingale residual process*

$$\hat{M}_i(t) = N_i(t) - \hat{\Lambda}_i(t) \tag{50}$$

When evaluated at the upper limit τ of the study, we have the *martingale residual*

$$\hat{M}_i = \hat{M}_i(\tau) = N_i(\tau) - \hat{\Lambda}_i(\tau) \tag{51}$$

Where it shows the residual between the observed and the estimated outcome by the model.

4.2.6 Model check and hypothesis testing

One of the assumptions made that needs to be fulfilled is the proportional hazards assumption. In the case that the hazard and covariable are linearly correlated in a similar way to other regression models. This can be checked visually by making sure that the hazard plot of the model is proportional to a satisfactory degree in comparison to other variables in the model. In other words, the assumption does not hold if the curves diverge or cross each other.

Alternatively, one can test the null hypothesis of non-proportionality for either each term or all the terms in the model, which is the method utilized by this thesis.

In order to test the null hypothesis of $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ where $\boldsymbol{\beta}_0$ is a known value, one applies one of the usual test statistics using the score function $\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta} logL(\boldsymbol{\beta})$ and the observed information matrix $I(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial\beta_h \partial\beta_j} logL(\boldsymbol{\beta})$ with three distributional and asymptotically similar test statistics:

- Likelihood ratio test statistics

$$\chi^2_{LR} = 2\left\{logL(\hat{\boldsymbol{\beta}}) - logL(\boldsymbol{\beta_0})\right\} \tag{52}$$

- Score test statistics

$$\chi^2_{SC} = \boldsymbol{U}(\boldsymbol{\beta_0})^T I(\boldsymbol{\beta_0})\boldsymbol{U}(\boldsymbol{\beta_0}) \tag{53}$$

- Wald test statistics

$$\chi^2_{W} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})^T I(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \tag{54}$$

With test statistics, we can utilize the common method of null-hypothesis testing for χ^2 distributed testing with p degree of freedom. As an example, if we use the Wald test statistics, we would use the following test settings:

$$Z_W = (\hat{\beta} - \beta_0) I(\hat{\beta})^{1/2} \tag{55}$$

In general, all three are roughly equivalent in their results and are mostly chosen based on preferences. These are used to test whether the current set of estimated parameters $\boldsymbol{\beta}^*$ is indeed reasonably justified as not being equal to zero.

4.2.7 Stratified Cox model

The regression model presented in the opening section describes a model where we assume that all the individuals share a common baseline hazard. However, for the most part, this is not the case, for example, where factors such as age are

important contributing factors to comorbidity in most health diagnoses. Due to varying baseline hazards, one can instead divide the individuals into groups or strata sharing the same relevant baseline hazard.

Assume that we have the same data set as the original Cox model with the population divided into k strata or groups, we define the regression model for individual i in strata s

$$\alpha(t|\boldsymbol{x}_i) = \alpha_{s0} r(\boldsymbol{\beta}, \boldsymbol{x}_I(t)) \tag{56}$$

Note that while the baseline hazard α_{s0} may vary, the covariates for each stratum remains constant. Additionally, the stratification also involves time dependency, it however per assumption of martingale that the information is based on past and not future information.

For the estimation of a strata model, we have the vector of $\boldsymbol{\beta}$

$$L(\boldsymbol{\beta}) = \prod_{s=1}^{k} \prod_{T_{sj}} \frac{r(\boldsymbol{\beta}, \boldsymbol{x}_{ij}(T_{sj}))}{\sum_{l \in R_{sj}} r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_l(T_{sj}))} \tag{57}$$

For $T_{s1}, T_{s2}, ...$ is the time of each event in strata s for the relevant risk set R_{sj} with the same properties that follow it as in the non-stratified estimation. Furthermore, martingale residual processes are adaptable to stratified as well, and the *Breslow estimator* is equivalently:

$$\hat{A}_{s0}(t) = \sum_{T_{sj}} \frac{1}{\sum_{l \in R_{sj}} r(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_l(T_{sj}))} \tag{58}$$

5 Model

The analysis in this paper is characterized by two components: the main survival analysis of the actual data set with the Cox model and the analysis of the simulated data set with the same method. Simulation and graphics presented in this thesis were made using RStudio version 4.2.2 with the packages *survival* for the calculation and data management.

For the first part, we make use of the real data set, with it only differing per definition. As a reminder of the three definitions used for the data set,

- ITT; We assume the individual has the same medication throughout the study, ignoring any gap periods, the simplest definition.

- OT; We emphasize the individual using the same drugs, as such any switches from the original drug or extended gap periods is considered censored beyond that time point.

- TV; Where the exposure is a time-dependent variable, such that each individual contributes to multiple drug types and exposure periods. Refer to **Chapter 3** for details.

For the OT and TV definitions, we assume any period where the gap period is longer than 15 days as censored or as **Non-exposed** period, respectively.

5.1 Non-parametric model

For the initial part of the project, we decided on a non-parametric model. Since we want to compare and contrast the three definitions, it's helpful to attain a quick overview regarding the overall effects on survival rate.

As such, by setting up a Cox model with no covariate and a time-dependent version by using Kaplan-meier estimation, we have a non-parametric survival function for the three definitions for each specific drug type. Its main purpose is to provide a general overview of the correlation and causation of the outcome and drug exposure period by just using the outcomes.

5.2 Cox model

The hazard model for the semi-parametric Cox model can be explicitly surmised as:

$$\alpha(t|x_i) = \alpha_{s0}(t)exp(Age_i + Sex_i + Hyper_i + MACE_i + Obesity_i)$$

Where α_0 is the baseline hazard for individual i for each strata of drug s.

The Cox model will be used to determine how the magnitude and confidence are affected by each definition. Elaborating on the covariate, we have the following factors with the implementation of strata for each drug category:

- **Age**; *continous integer*, as of the start of the study
- **Sex**; *binary*, 1 = Male, 2 = Female
- **MACE**; *binary*, 0 = No prior history of MACE
- **Hypertension**; *binary*, 0 = No history of hypertension
- **Obesity**; *binary*, 0 = Not obese on start of study.
- **Code**; *Categorical* = {1,2,3,4,5}, Differing drug regime

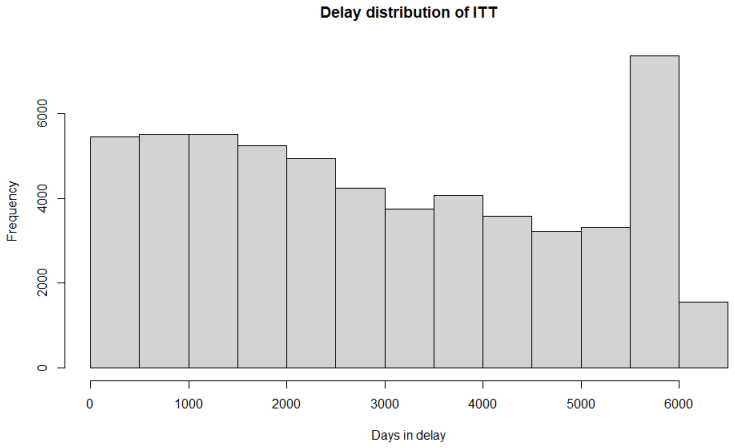

Figure 8: Distribution of days spent in the study for all three definitions, notice how ITT follow a somewhat uniform structure.

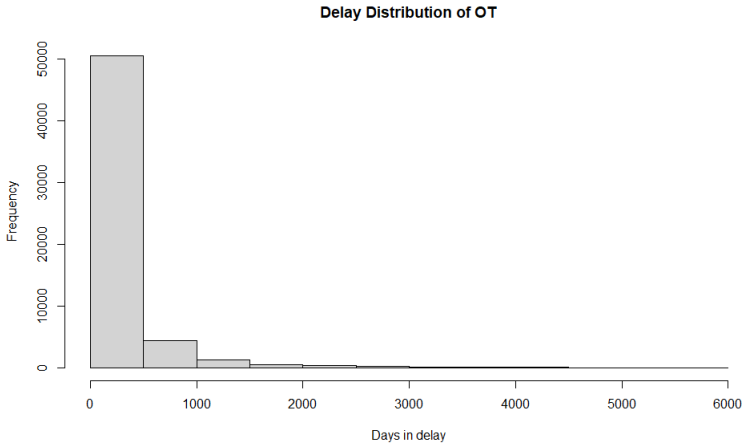

Figure 9: Total delay in the initial exposure period by OT definition, in other words, how long each is in a given treatment before switching. Notice the exponential distribution-like structure; when interpreted alongside Figure 8 it implies that many participants don't stay too long on a given medication and switch often when comparing them in the total amount of delay.

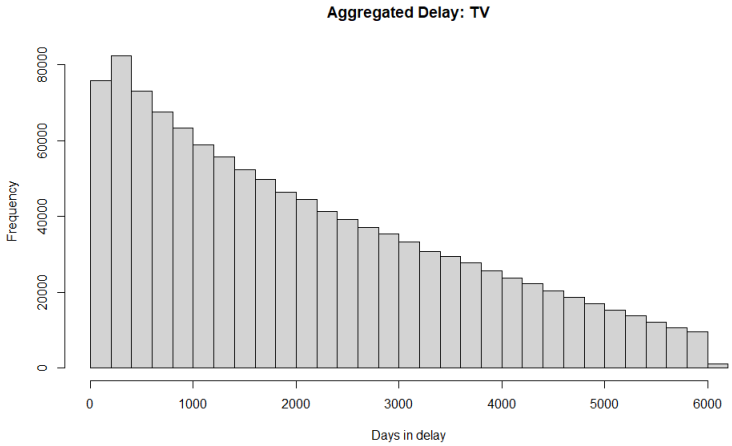

Figure 10: The aggregated delay of all the exposure periods, TV differs due to how the exposures are assigned, but note that most uninterrupted exposure periods lie on the shorter side.

This selection of variables was chosen due to their perceived relatedness to cardiovascular conditions, as the expectation was that a small selection of significant covariates would be prudent to display any potential differences in the three definitions.

We additionally let **Code** be used to stratify the model where $s = 1, 2, 3, 4, 5$ is the five drug category on the baseline hazard function α_{s0}, allowing the variable to contribute without making the model too complicated. As such, in this model, we assume that each drug has an underlying hazard rate similar to the other.

6 Simulation

After we have completed a Cox model of the data set, we now wish to test the definition in a controlled setting, putting the three definitions into extreme cases.

6.1 A simulated data set

Even with a Cox model fitted for each definition, using only the actual data set does not allow for a satisifiable statistical conclusion to be reached due to missing many important factors and assumptions for the remaining factors as well.

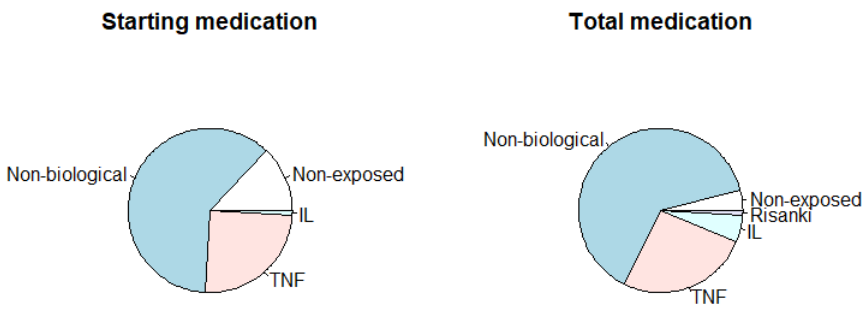

Figure 11: Pie chart for the fraction of exposure periods in the data set, the starting one displaying the initial medication relevant for ITT and the total amount of uninterrupted exposure periods as defined by TV for the duration of the study. Risankizumab was not available at a large amount at the start of the study, hence its exclusion in the left pie chart.

Thus, the intention is to make use of a new set of data based superficially on the real data set, which will allow us to further stress test the three definitions in a more controlled manner conducive to more mathematical conclusions. Also of interest is whether the definition handles data sets with skewed results, such as higher intensity of outcomes or disruptions in the individual's treatment plan.

6.2 Generating a new set

As the simulated data is needed to be ostensibly based on the real data set, we put forth an algorithm for how we intend to simulate this new data set.

6.2.1 Participants

The initial set of data could be generated however one wants; as such, the initial set of participants is generated proportionally in regards to the actual data set, as shown by Figure 3.1, which in turn allows for a quick setup of somewhat equally distributed variables for the participant and the actual data set.

6.2.2 Event time

With the main set of participants, the following event times are generated based on the results of the Cox model utilizing the inverse survival function defined as follows:

$$t = h^{-1}(\frac{log(U)}{exp(x * \hat{\boldsymbol{\beta}})}) \tag{59}$$

Where x is the covariate of an individual, $U \sim Unif(0.1)$ and $\hat{\boldsymbol{\beta}}$ is the coefficient attained by the Cox model. Additionally, we set the baseline hazard to $h_0(t) \sim Wei(0.5, 4.5)$ as the Weibull distribution is commonly used in survival analysis, thus the explicit inverse survival becomes:

$$t = \left(-\frac{log(U)}{\lambda exp(x * \hat{\boldsymbol{\beta}})} \right)^{\frac{1}{\rho}} \tag{60}$$

Where it is multiplied by 12 000 such that the mean becomes more in line with the distribution in Figure 8 for ITT using that definition as a base and reference due to it retaining most of the total exposure period in the actual data set.

6.2.3 Censoring

For the censoring time, based on the results of the number of days each participant has spent in the study, we use Figure 8 with its uniform distribution as a base and set the censoring time as $C \sim Unif(0, 6000)$.

Whichever the censoring time or the event time is lowest, we consider it as observed; in other words, if the censoring time is lower than the event time, we consider that event censored, and vice versa for the event time.

6.2.4 DDD

The DDD were chosen from a sample data set of the actual data, as an exploratory analysis of the real data showed that the DDD amounts were to be pre-determined based on the medication and not generated by chance or random number.

6.2.5 Switch Time

As noted by the actual data set, such as Figure 9 and comparing it to Figure 8, not many linger too long on a single medication, with some being quite liberal with it and often switching from it after some period. We call this period switch time, the duration in which the individual is on one drug, and based on the same figure, we determined it as an exponential distribution, more explicitly:

$$t_{switch} \sim Exp(0.0006)$$

Also observed was that the DDD value was constant throughout that period, only changing when the patient received a new drug. So for each switch time, we fill each entry with the drug, date, and DDD with a randomized gap interval

between each entry using a combination of exponential and poisson distributions with a mean of 10 time units. Explicitly we use the formula:

$$G = Pois(20) - Exp(0.1)$$

Where G is the gap between the expected DDD periods.

This is repeated until we reach the end of the switch time, at which point a new switch time is generated, or if we reach the censoring or event time, at which point we repeat the generation for the next participant.

Which drug is assigned is based on the same proportionality of drug usage present in the real data set, which is illustrated in Figure 11. This, in practice, means we can expect a sort of proportionality in play between the drug and the survival rate on the simulated data set, though it likely won't capture the true interplay present in the real data set.

6.2.6 Modelling unpredictability

To add further stress to comparing the definitions, we also wish to inject some unpredictability into select data sets. An example of such is if the outcome is of higher or lower intensity, whether the individual misses their prescription dates, and other elements of that nature.

As such, the two main ways we have chosen to mutate the results are the intensity of outcomes, abstracted as shorter event times, and adherence to the treatment plan, which is abstracted as increases or decreases in the gap periods.

When we have generated the new data set, we put it through the same data processing as the actual dataset and compare the Non-parametric and Cox models. A overview of the generation process can be viewed from **figure 12**

6.3 Test models

We generate three distinct models to contrast them: the control, intensive, and lazy models.

6.3.1 Control model

The control model, as the name suggests, is the basis on which we test and also base our interpretation in comparison to the two other models. As such, any unpredictability is set to default; in other words, there are no changes in the number of events or the gap period.

This model also serves to compare the simulated data to the actual data set to see how well the data was simulated without any modifications.

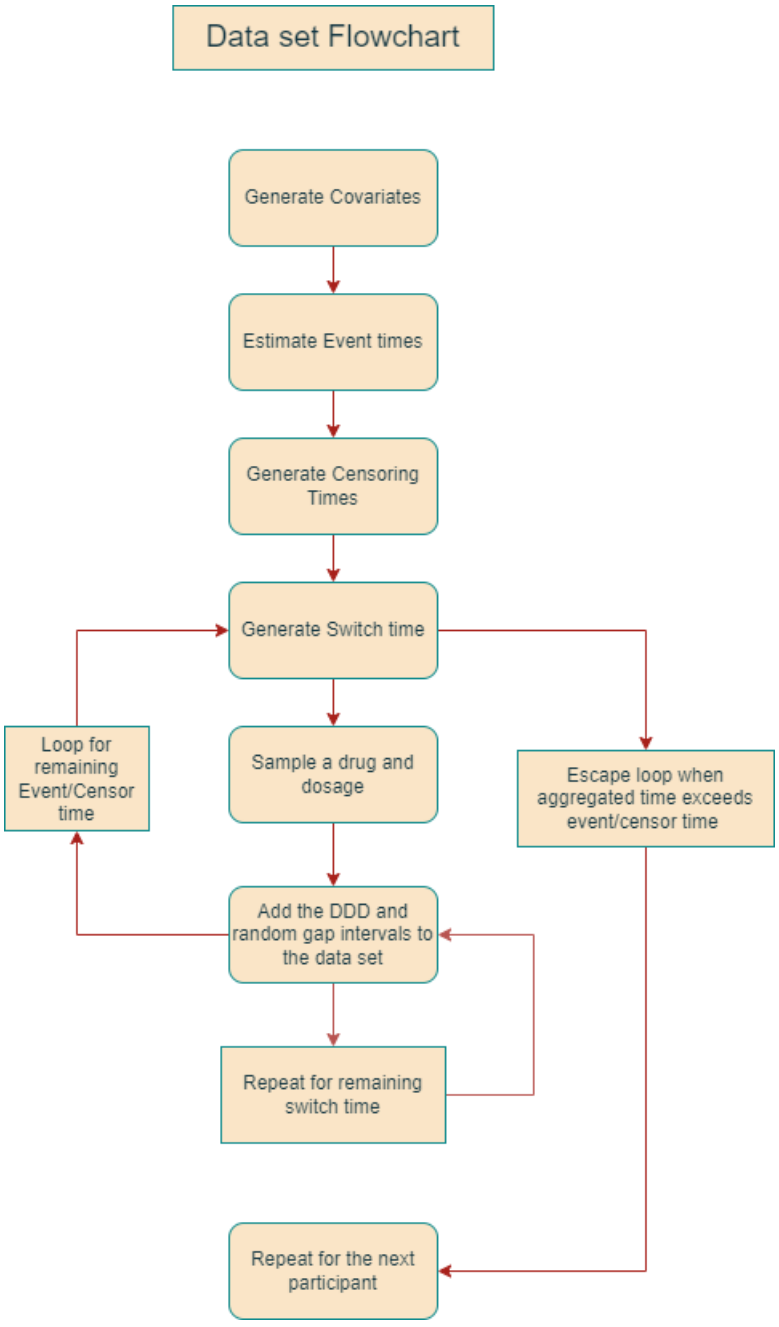

Figure 12: Flowchart for generating a data set for each participant

6.3.2 Intensive model

In the intensive model, we want to simulate a data set where the event is more widespread and common; this is done by decreasing the event time across the board, such that more outcomes bypass the censoring step. In effect, we can expect the new data set to contain more and earlier events compared to the control model. In practice the time to events becomes:

$$t = \frac{t_{SurvEvent}}{I}$$

Where we $t_{SurvEvent}$ refers to (60) and I is the intensity parameter we set.

The context behind the event time might be strange considering how we defined the event or censor time, but we digress and leave it to the discussion section.

6.3.3 Lazy model

The Lazy model is made to model a data set where the individual has low adherence such as not following the prescribed schedule for dosage or refill. In effect, this is abstracted as the gap between each refill increasing as lack of adherence is interpreted as a lower dosage or less frequent refills. Explicitly we set that:

$$G = \frac{Pois(20)}{A} - Exp(0.1) * A$$

Where A denotes the parameter for laziness.

While this method ignores some other possible sources of adherence, such as taking a large dosage or refilling too early, due to the way we defined gap periods, this isn't represented in the data processing.

7 Results

In this section, we aggregate the results from both the actual and simulated data sets contained within each section.

7.1 Main data set

This section concerns the model based on the data set from Social Styrelsen.

7.1.1 Non-parametric measure

In the initial overview, we examine how the definitions compare the hazard rate given only the outcomes and their time delay.

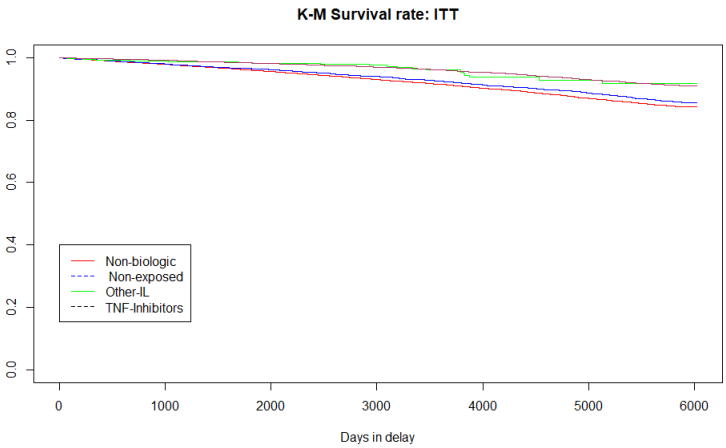

Figure 13: Survival plot for ITT for the real data set, displaying low hazard rate.

	Age	Sex	Hypertension	Prior Mace	Obesity
ITT	0.06	-0.48	0.37	0.79	0.19*
OT	0.059	-0.34	0.48	1.15	0.11*
TV	0.069	-0.48	0.37	0.70	0.19*

Table 1: The fitted coefficients of the Cox model, where * denote a particular parameter with less than ideal confidence ($p > 2e^{16}$)

ITT offered the easiest way to aggregate the data, and it displays in Figure 13 stable and low hazard rate. It detected 3815 outcomes, which is indeed what the data set itself recorded.

OTdisplays some jagged curves on Figure 14 which stem from the reduced outcomes assigned in the data set per its definition, from 308 events compared to 3815 in the actual data set. The survival rate surprisingly aligns with the results suggested by ITT despite the lower count of outcomes, likely due to proportionally less exposure time assigned in the definition.

The TV method offers the most radical difference of the three definitions in Figure 15, inflating the hazard for the **Non-exposed** category, while the other four still align towards the result by ITT. It could be explained since all four drug categories also contribute to **Non-exposed** which means that the category has more opportunity to inflate its hazard rate.

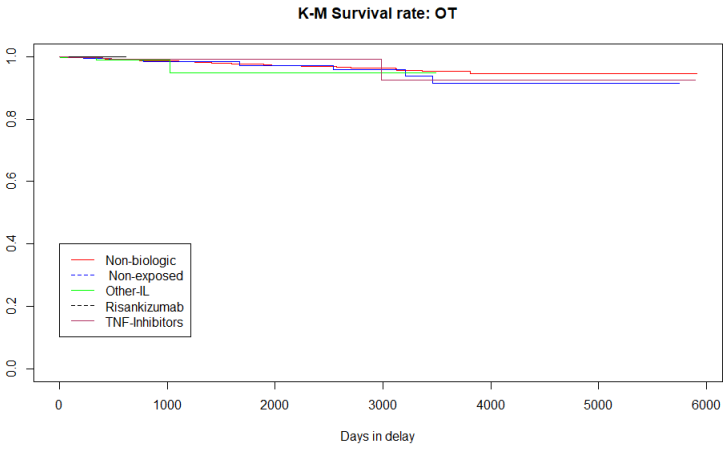

Figure 14: OT survival plot set in the real data set, where we notice more jagged curves suggesting less statistical power due to fewer observations. Though the survival rate is in line with ITT.

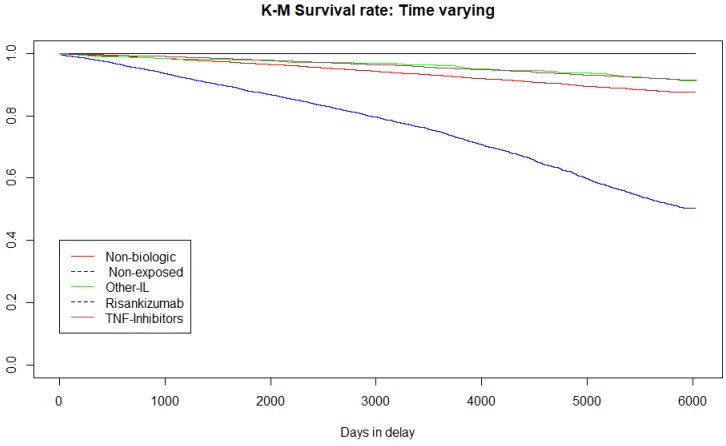

Figure 15: Survival plot for TV method, notice that there appears to be a "inflated" risk associated with **Non-exposed** category.

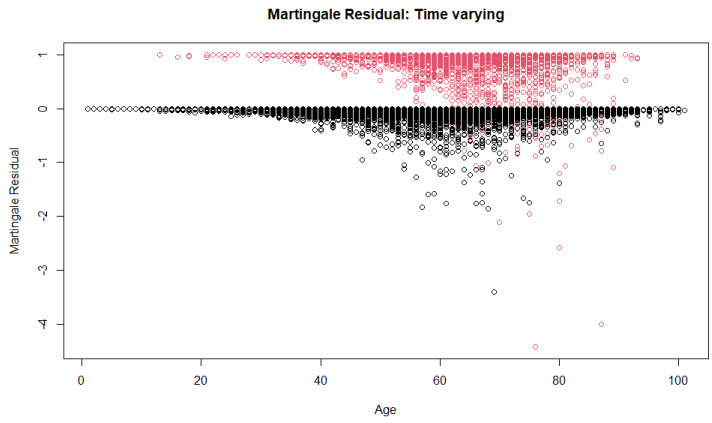

Figure 16: Martingale residual of the TV method: red dots indicate outcomes, and if any dots bleed below 0, we have a miss-classification at hand.

7.1.2 Cox Model

The Cox model displays the magnitudes for the parameters of interest, with positive values equating to correlating effects for risk factors in contrast to negative values presented in table 1.

The Cox Model variant of the definitions displays the same proportional difference in its covariates; ITT and TV lie in the same ballpark in regards to the magnitude of their coefficients while also displaying the same level of confidence on each parameter. OT was oddly quite separate from its two counterparts, with its increased magnitude of Hypertension and Prior Mace being most noticeable, likely stemming from the fact that we have a shorter exposure period as a whole, so the hazard rate likely increases as a consequence. Suggesting that the three definitions are not equivalent in their results.

Obesity in general was considered low-confidence in its correlation for all three definitions, and given that obesity is often miss-classified and something of a definition question, it's reasonable to assume that the state of obesity is of questionable value. Of course, it could be regarding differing degrees of obesity, but that's speculation.

Checking the diagnostic on Table 2, it seems that the assumption of proportional hazard seems to hold, if barely only for ITT, The Martingale residual in Figure 25 shows that the outcomes don't diverge too far from its expected outcome. We also note that there appears to be some misclassification on the TV residual plot in Figure 16. The following diagnostic plots are referred to

p	Age	Sex	Hypert	MACE	Obes	Global
ITT	0.838	0.028	0.228	0.053	0,821	0.114
OT	0.213	0.277	0.941	0.006	0.274	0.040
TV	0.088	0.117	0.253	0.021	0.843	0.027

Table 2: P-table for proportional hazard, where's the null is non-proportional hazard

	Age	Sex	Hypertension	Prior Mace	Obesity
ITT	0.058	-0.66	0.39	-0.23*	0.62
OT	NA	NA	NA	NA	NA
TV	0.058	-0.66	0.38	-0.22*	0.61

Table 3: The fitted coefficients of the Cox model for the Control model, NA were added to the OT row due to lack of viable model for the simulated data set.

henceforth in the Graphical Plot Section on pages 51 to 55.

In short, ITT and TV method seemed to be in practice equivalent in terms of Cox parameters, while OT emphasizes the hazard rate, increasing it for the same allotted exposure period, though the diagnostic property makes this assumption questionable. In contrast to the results suggested by the non-parametric model, where the survival rates of ITT and OT were equivalent, the TV method was the outlier with its inflated risk in the **Non-exposed** category under the context of a non-parametric model.

7.2 Simulated data set

In this part, we analyze the simulated data set using the same data processing process used for the actual data set. All three present models were generated with the same 10,000 participants, with only the event time and adherence being modified.

7.2.1 Control model

The control model were simulated with no modification for adherence or intensity, yielding us 551(5%) observed outcomes in line with the main data set.

The control model displayed notable divergence from the actual data set that these generated data were based on; ITT as an example, didn't capture the differing survival rate for each drug, nor did it display the same general survival rate in the Non-parametric model. It's notable that the Cox parameter is quite different compared to the ITT in the main data set, as the importance of obesity and prior MACE were switched, with the magnitude being radically different as well.

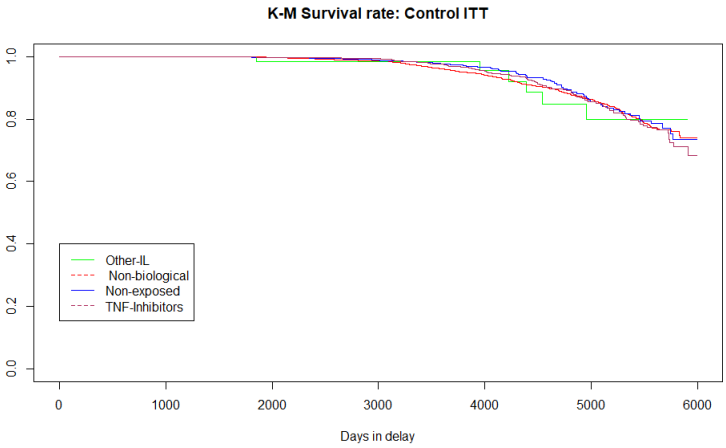

Figure 17: Survival plot for ITT in the control model, where's a increased hazard rate is induced by the simulation

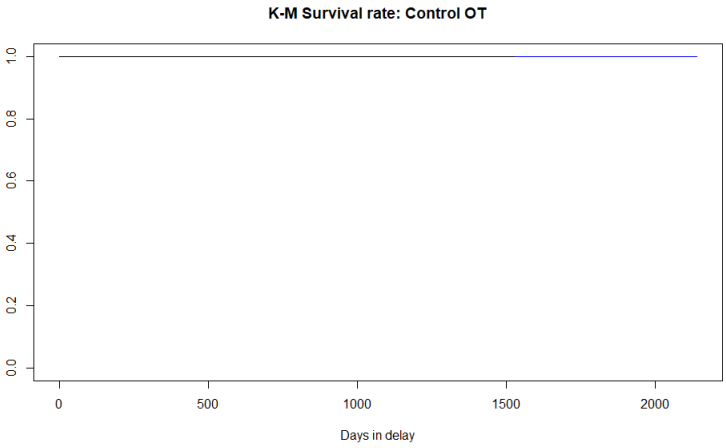

Figure 18: Survival plot for OT, where by its definition didn't detect any outcomes

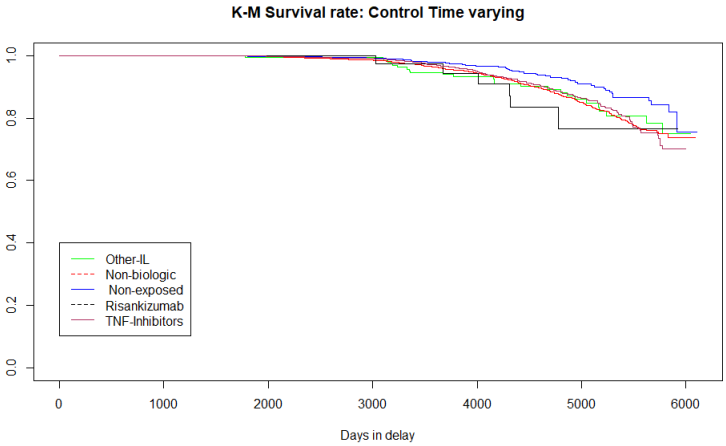

Figure 19: TV in this simulated data set retains the proportionality defined in the simulated steps, showing that given how each outcomes is discovered/written down determines the results

OT model didn't appreciate the new model, because the drastic amount of less observation lowered the already low number of events observed. Even at a level that makes an appreciable Cox model impossible, in comparison, the real data set (n=57764) had 308 observed outcomes in the OT definition. The Control model had no events observed.

Comparing TV and the ITT definition in the control model, we note again that the two definitions seem to be equivalent in the results of the Non-parametric model, suggesting great divergence from the base data set, though it likely stems from the way the data were generated with most of the outcome occurring during the expected DDD period. ITT and TV are also equivalent in the Cox model, which is in line with the base data set.

The diagnostic plot seemed to be agreeable with the simulated data, as the proportional hazard assumption in table 6 holds for both definitions in contrast to the real data set, and the Martingale plot with Figures 27 and 28 doesn't show too much bleeding of the outcome, confirming good prediction and linearity of the model.

7.2.2 Intensity model

In the intensity model, the event time for any observations was halved, which consequently increased the number of outcomes to 4001 (40%) marking a sharp increase in outcomes.

	Age	Sex	Hypertension	Prior Mace	Obesity
ITT	0.059	-0.47	0.38	0.20 *	0.62
OT	NA	NA	NA	NA	NA
TV	0.077	-0.67	0.50	0.39	0.82

Table 4: Cox Coefficients for intensity model: we note the shifting magnitudes for Hypertension compared to the results of Control model, refer to Figure 3. In addition we note that the parameters for ITT and TV definitions now differs in contrast to the Control model. While OT did observe some events, they were not enough for a satisfactory Cox model hence the NA on its row.

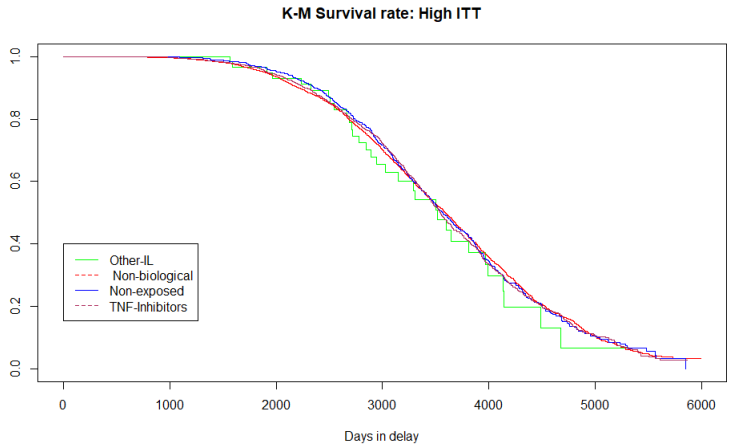

Figure 20: High intensity survival plot for ITT where the proportionality of the drug types is retained.

The ITT definition surprisingly handled the increased outcomes quite well, owing to its admittedly simple data processing. The survival rate in the Non-parametric model showed a rather decreased survival rate, which is to be expected from the modification, but the proportionality of the drug types is still present. The Cox model, in turn, has more positive coefficients to make up for the decreased survival rate; most notable is that MACE is now positive again and Age is more of a contributing factor.

OT even with the increased observed outcomes, still didn't capture enough observations to make for a passable Cox model or anything at all. Of the 4001 outcomes, only two passed the process, which states more the fact that adherence is more important to this definition than the volume of outcomes.

TV method showed some concerning results; instead of continuing the trend

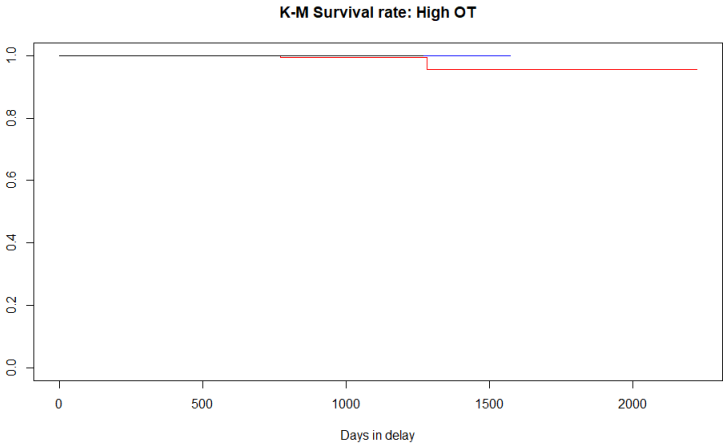

Figure 21: Even with the heightened intensity, OT didn't detect enough outcomes to make for a usable model.

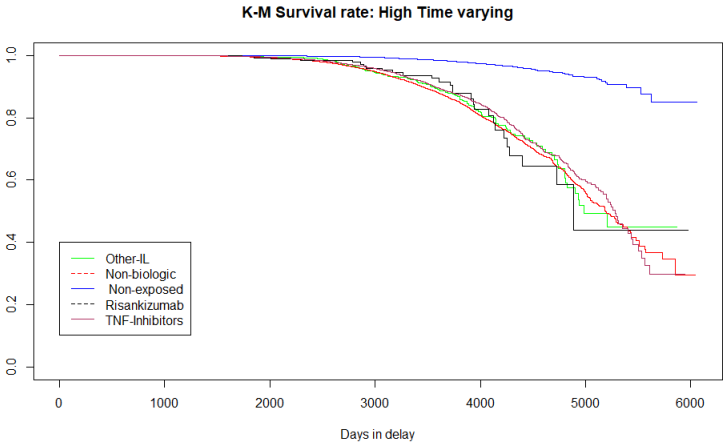

Figure 22: Interestingly, in the intensity model, the drug category **Non-exposed** is now underrepresented in the survival rate compared to both Figure 14 and 17.

	Age	Sex	Hypertension	Prior Mace	Obesity
ITT	0.057	-0.64	0.54	0.05 *	0.81
OT	NA	NA	NA	NA	NA
TV	0.057	-0.63	0.55	-0.077 *	0.88

Table 5: Cox parameters for Lazy model, here we note that ITT and TV differs in the Cox parameters induced by the participants low adherence. OT were again omitted due to likely producing no better results than the Control model.

of being roughly equivalent to ITT we now note extreme divergence.

The non-parametric model showed a higher survival rate than ITT with **Non-exposed**, being the least likely to experience an event, which is a severe departure from the norm produced by the base data set where the **Non-exposed** category was the most inflated. In addition, the proportionality between drugs that was present in Control model is not retained in this model, and the general survival rate is much higher in comparison to the ITT in the high intensity model.

The Cox model did not appreciate the increased outcomes, diverging completely with its results from the expected norm by ITT, with it now stating that all the parameters are significant.

The diagnostic shows some strong miss-classification with its Martingale in figure 29 for ITT, denoting non-linearity. Its counterpart, TV showed no such degree of miss-classification, though interestingly, proportional hazard is retained in ITT while it doesn't hold for TV.

7.2.3 Lazy model

In the lazy model, we have the same event time, but the expected gap between refills is drastically increased with the same amount of observed outcomes (563), in line with the control data.

ITT showed no difference from the control model, as the ITT model only concerned itself with the starting drug and the eventual censoring or event time.

OT probably won't make the cut; if it stuttered on the control model, a model with a noted lack of adherence likely won't do much good. If the adherence were instead increased, this likely would converge to the ITT definition suggested by Figure 14.

TV as the only model to be sensibly impacted by the adherence factor, the Non-parametric model showed a lessened survival rate, with the **Non-exposed** again having the highest survival rate, likely due to the increased amounts of "harmless" **Non-exposed** periods.

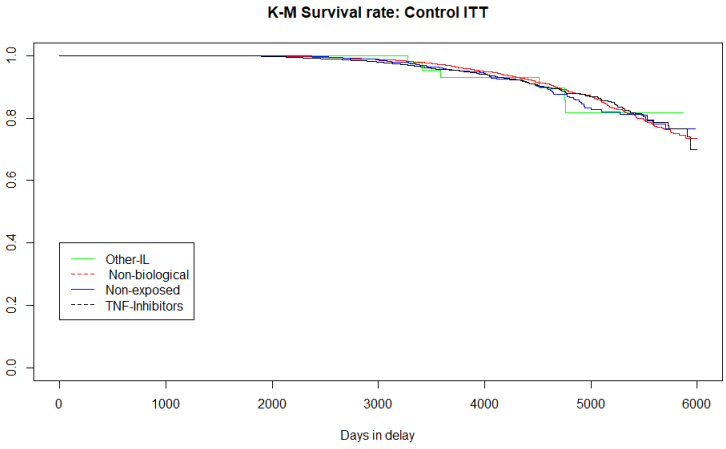

Figure 23: *ITT* for low adherence, where it as expected is unaffected by the adherence level

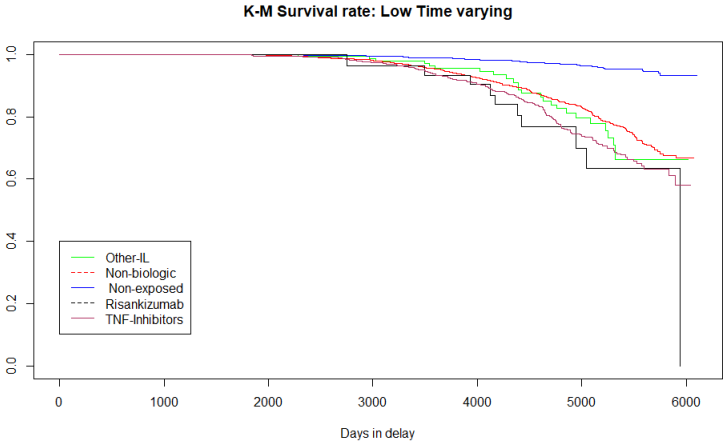

Figure 24: We notice again that the Non-exposed category is again underrepresented for the *TV* method, and there appears to be a drop in one of the curves with particular jaggedness. Suggesting a low amount of events related to that drug

With the exception of Prior MACE, ITT and TV shared the same magnitude and confidence for the same parameters, though this does display some inequalities present between the definitions.

The diagnostic showed that the proportional hazard is satisfied for both ITT and TV by table 8. The martingale plot for TV in Figure 32 showed some notable bleeding for the outcomes, implying some non-linearity in the model.

In short, it appears that inducing some unusual conditions in the simulated data makes the definition diverge in its results; specifically, it appears that while TV and ITT remain somewhat equivalent in the Control model, any modification removed that common element for both higher outcome and low adherence parameters. Implying that for conditions where we can expect some extreme non-standard conditions, the three definitions start to differ severely.

8 Discussion

8.1 The base data set

Some of the entries in the data set had individuals experiencing an event before any drugs were dispensed. This mostly concerned ITT given the definition of whether one starts at the beginning of the study entry or the start of the actual treatment. In this case, they were left out as censored outcomes, meaning they didn't contribute to the model.

The namesake of the drug, Risankizumab, was not available in large numbers during the early days of the study, hence why it only showed up as a switched drug and not as an initial medication. This has no large mathematical bearing on the thesis but has an interesting effect on the definition used, as the two definitions, ITT and OT, only care about the initial treatment plan, which means that one of the major drugs used is not represented. Again, this likely has a clinical interest when one uses ITT as it uses the whole exposure period and most likely plays a role in the event of an outcome.

Though, in regards to the OT part, there was one participant assigned as Risankizumab in contrast to ITT, this is likely just a quirk with the original data set as we detected too few participants for it to be noticeable.

Figure 15 displayed a notable inflated risk associated with **Non-exposed** category; again, as mentioned in the results themselves, it could stem from the fact that the four main categories contribute to Non-exposed with its gap period, so the dump category has more opportunities for more outcomes. Another way to see this is that most outcomes are assigned when they're detected and not the actual date they occurred. As such, it could be that most of the symptoms

are detected at a later date, such as the refill date, and then the outcome would be assigned as **Non-exposed** exposure period.

8.2 Cox model

Some notable steps were taken when fitting the Cox model; the most apparent is the set of variables and stratifying the model by drug type. It was decided to do so since having the categorical parameters of each drug type would complicate the model to an unreasonable degree.

The current set of variables was chosen more as a set of variables that could easily be interpreted for the effects each definition has. Hence, there was no extensive process for choosing a set of parameters, as the main purpose was to find comparable parameters and not an exhaustive examination of the efficacy of any psoriasis medication. The fact that one of the parameters, obesity, was prone to miss-classifications is a benefit to us as a way to test how the definition handles less reliable parameters.

In addition, no attempt was made at finding any interaction between the variables, as again, we were interested in comparative parameters and not the actual true parameter for the outcomes.

Some interesting thoughts about a specific definition, OT, were that, given the treatment or set-up of the data set, it could either return a data set usable for survival analysis or, more unhelpful, nothing usable. For the actual data set, we were fortunate enough to have anything, but the loss of observed outcomes (308 against 3815) was almost not enough, and we still would prefer more observed outcomes. If the data were any more diluted, smaller, rarer, or in any combination, we wouldn't have any usable data, as demonstrated by the simulated data set.

Interesting enough, even the drastic reduction of observed outcomes still returned the same level of survival rate compared to ITT, likely due to the proportionally lowered exposure periods in general for OT. It can be interpreted that OT can return a sensible survival rate if given a reasonable amount of data.

8.3 Simulated data

While the intent of this thesis was to measure the difference in hazard rate between the three definitions, comparing those in the same equally distributed data set should be sound reasoning. Though the simulated data set turned out to be diverging from the real data set by the start, some analytic value could still be had by examining the results internally between the three test models.

The varying results shown by the TV method showed a lot of variance when given a data set with extreme conditions. While the divergence in the Control

model from the real data set likely stems from the way the data were generated, it still shows that how the data are laid out or collected affects the end model. This goes further internally in the simulated data, which demonstrates that TV method starts displaying wildly different results compared to ITT when applied to modified data.

This then led to some observations regarding the TV method: given preset parameters by the user, it can drastically alter the results. Indeed, if we, for example, increase the gap delay in the two modified simulated models of **Non-exposed** category, we could expect a less inflated survival rate due to creating fewer "harmless" exposure periods. This also extends to the actual data set, where such a change could result in us assigning the outcome to another category that's not **Non-exposed** hence its "proper" category.

Conversely, how the data set is collected could also matter, since if most of the outcomes occur during the gap period, then it is expected that we will have an inflation of the hazard rate on the **Non-exposed** which we saw in the real data set. It probably is in one's interest to introduce another "dump" category for any potential gap in the refill period to at least avoid these kinds of results in the future or at least differentiate between actual **Non-exposed** and gap periods.

This again relates to how TV performed in the actual data set, where Non-exposed were inflated and the intensity and lazy models were deflated. This could be interpreted as meaning that in a controlled environment where we know the true dates, **Non-exposed** period is de-emphasized while the data set with no known true date emphasizes the risk associated with the **Non-exposed** period, if, of course, the simulated data is as close to life as possible.

The results of the OT on the simulated data are unfortunate and could be seen as a waste of analysis; however, they serve some analytical value. It shows that OT doesn't detect any outcomes if the adherence is too low, depending on the user's definition of the course. It also shows that even if the data set has a high hazard rate, OT wouldn't still detect if the adherence is still too low, arguing that OT needs some pre-data processing or exploration before being considered.

When the event and censoring time were simulated, we censored all event time that occurred after the censoring time. This is strange since this implies we assume everyone in the study will eventually experience MACE, and we haven't observed it because it hasn't happened yet. In the context of survival analysis, this is fine, though that approach could be a factor in why we had a dip in the survival rate at the end of the study period on the simulated data set.

It should be stressed that the simulated data set is made with assumptions and caveats; despite the attempt to base the simulated data on the real data

set, digressions and shortcuts were taken to explain why the simulated data set is based on and is not the real data set. A factor that wasn't used were any potential hospital or health care visits; it doesn't affect the thesis in any major way, but one of the qualifications used for the definition is that healthcare visits are taken into account.

When generating the covariates for the participant, a simple spread done by uniform distribution was used to assign the covariates. This could create some entries that could be considered nonsensical, such as a 1-year-old having prior health complications from cardiovascular disease, which likely don't happen in reality. Of course, given that we based this on the real data set, this shouldn't really affect the general mean of the model, but it still retains some compromise made for abstraction over realism.

One of the assumptions made is that the prior drug used in an individual's prescription history does not have a lingering half-life effect; in other words, we assume that the prior medication has no bearing on the current survival rate. In extension, due to the Cox model chosen, the underlying base hazards inherent in each drug were largely omitted for a more readable model with it being stratified by the drug category, hence why the actual choice of medication actually doesn't have an effect on the survival rate, though the simulated data represented this as proportionally randomly divvying the drug to each period, which is fine on a macro level but does not represent the actual drug usage or its efficacy.

Another part that's suspected but also outside the scope is the interaction between exposure periods. It stands to reason that combining some of the drug in a certain way after a certain period could be a notable factor in the event of an outcome. Given that the Cox model was likely not catching this interaction, the simulated data also omitted this interaction, and realizing this concept on a simulated data set is a far deeper task that extends the scope of this work. This also concerns the Cox model of the real data set, as no thought was given to prior interactions between drug exposures, lending some digression regarding the precision of the thesis.

8.4 Misc

Some thought about a specific definition in general, ITT, is that it doesn't use the adherence of the participant as a factor to be concerned about. Indeed, given the way we processed the data set, ITT doesn't care how diligent the participant is. Of course, adherence does likely have an effect on the definition as one needs to use the drug to be able to experience the side effect, but during analysis that factor is missed more as a preventative measure than a direct correlation. Which likely explains its sturdiness in the simulated data set.

For the applicability of any extreme cases, while one hopefully won't find a situation where half of the participant experiences a side-effect, though it could be useful for another data set with differing context, the other model for laziness is definitely a more real and applicable situation one could find itself in.

9 Conclusion

Concluding the study, we have found that in terms of discussing the overall hazard of the models, the three definitions had wildly differing results, which seemed to stem from how the data were collected, processed, or interpreted. ITT was the definition that was the least volatile, which might be due to its simplicity; OT however needed some pre-data processing as if it's given a data set that's small, diluted, applying too much of a gap, or the participant has low adherence, it has low statistical power, but it can be expected to converge towards ITT and is a better fit for specific findings. *Time-varying* method shows stability when the data is relatively reasonable in line with ITT but when it's applied to a data set that's extreme in one way or another, it can produce wildly differing results. It's also noted that setting the parameters for *Time-varying* in regards to handling gap delays can adversely affect the end results, emphasizing the importance of parameters set by the user subjectively and not by a set standard, and that given that the data shows the true date of its event, it can either overstate or downplay the associated survival risk with TV. The simulation, while not true to the actual data set, does provide some insight into the interplay between the definitions when viewed in a vacuum and has lent itself to some observation displaying that for some extreme conditions, the Cox model of ITT and TV starts diverging in contrast to the stable results displayed on the real and controlled data set.

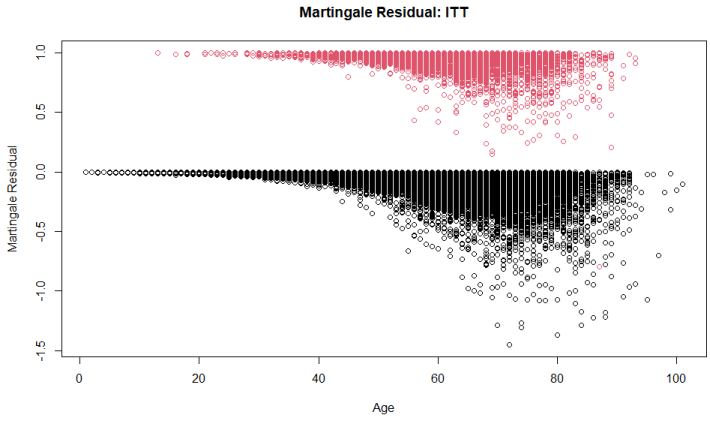

Figure 25: Base Data set, the following Martingale plots displays good diagnostic properties. They're for reference and posterity's sake, and they're presented as is.

10 Graphical plots

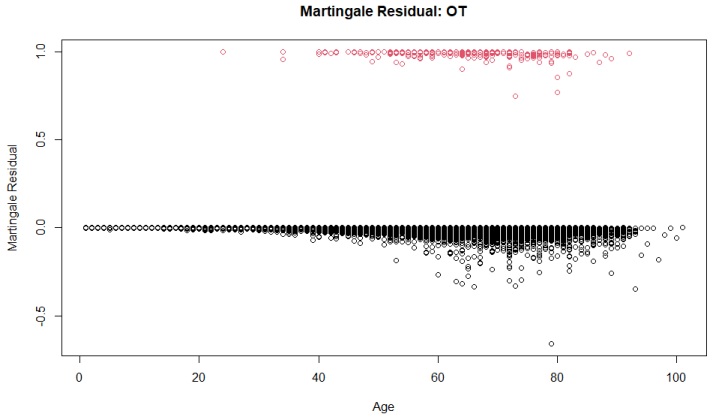

Figure 26: Base Data set

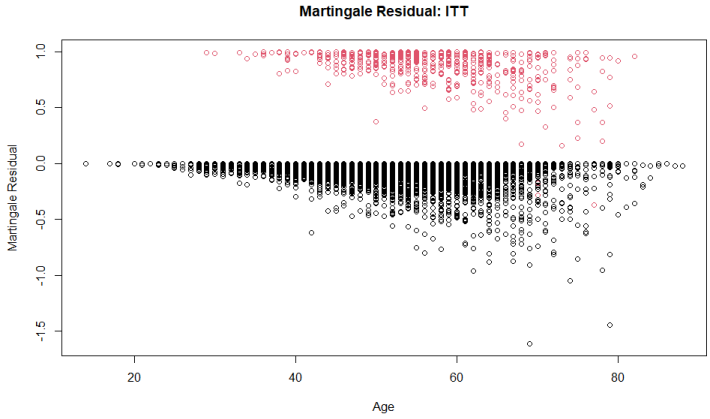

Figure 27: Control Data set

Figure 28: Control Data set

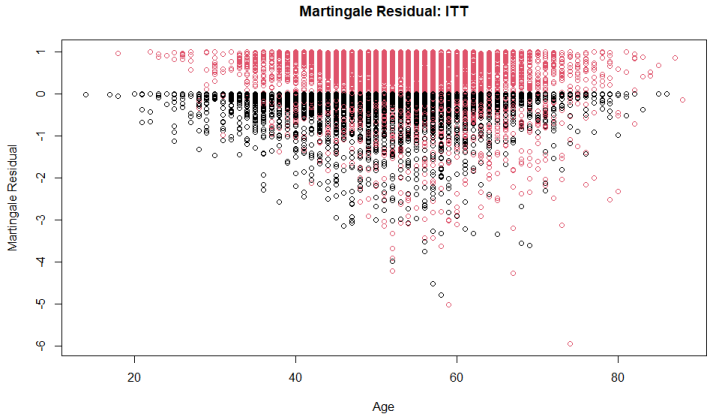

Figure 29: Intensity model Data set

Figure 30: Intensity model Data set

Figure 31: Lazy model Data set

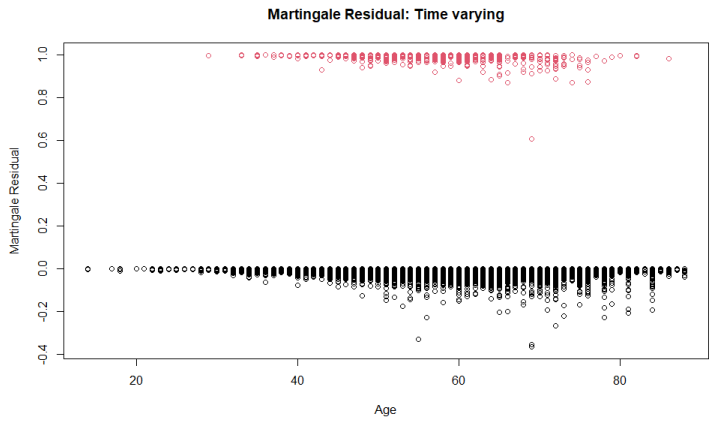

Figure 32: Lazy model Data set

11 Appendix

11.1 Proportional hazard values

p	Age	Sex	Hypert	MACE	Obes	Global
ITT	0.18	0.50	0.31	0.97	0.78	0.64
Time-varying	0.23	0.61	0.23	0.94	0.73	0.65

Table 6: P-table of Control model for non-proportionality, where we see the assumption of proportional hazards holds for the parameters and the model itself.

p	Age	Sex	Hypert	MACE	Obes	Global
ITT	0.61	0.86	0.92	0.32	0.75	0.93
Time-varying	0.01	0.009	0.23	0.036	0.96	0.004

Table 7: P-table of Intensity model for non-proportionality, we note that TV definition don't hold for proportional hazard assumption.

p	Age	Sex	Hypert	MACE	Obes	Global
ITT	0.58	0.77	0.22	0.66	0.28	0.63
Time-varying	0.55	0.74	0.36	0.26	0.38	0.64

Table 8: P-table of Lazy model for non-proportionality, where we again note good proportional hazard properties.

11.2 Predictable variation process

Given a time-continous martingale M, the predictable variation process is defined as:

$$< M > (t) = \sum_{k=1}^{n} Var(\nabla M_k | F_{k-1}) \tag{61}$$

In other words, it's defined as the sum of Martingale differences in its conditional variance.

11.3 Martingale central limit theorem

Let

$$M^{(n)}(t) = N^{(n)}(t) - \int_0^t \lambda^{(u)} du; \quad n = 1, 2, ...$$

where it's a sequence of vector valued counting process martingales denoted by $k^{(n)}$ for the dimension of $M^{(n)}$, then for each n we introduce p **x** k_n matrix $H^{(n)}(t)$ for the predictable processes $H_{hj}^{(n)}(t), \quad h = 1, .., p, \;\; j = 1, ..., k_n$ Consider the limiting behavior of the stochastic integral.

$$\int_0^t H^{(n)}(u) dM^{(n)}(u); \quad n = 1, 2, ...$$

In focus, we prove that for the condition for this sequence to converge to p-variate mean 0 gaussian distribution. Let's denote a p-variate Gaussian martingale as $U(t)$, its covariance is determined by the function $V(t) = E\{U(t)U(t)^T\}$. A continous determistic p**x**p matrix-valued function that's zero at time zero, and with positive increments $V(t) - V(s); t > s$

To have the sequence of vector valued counting process martingales converge, it's required that the predictable variation process converge in probability to the covariance function of the limiting Gaussian martingale; this takes the form

$$\int_0^t H^{(n)}(u) diag\{\boldsymbol{\lambda}^{(n)}(u) du\} H^{(n)}(u)^T \xrightarrow{P} V(t); \quad n \to \infty \tag{62}$$

For all $[t \in [0, \tau]$ where $\boldsymbol{\lambda}$ is the same intensity process offered at section 4, additionally another requirement states that the sample paths remain continous in the limits. Which can be stated as

$$\sum_{j=1}^{k_n} \int_0^t (H_{hj}^{(n)}(u))^2 I\{|H_{hj}^{(n)}(u)| > \epsilon\} \lambda_j(u) du \xrightarrow{P} 0 \tag{63}$$

For all $t \in [0, \tau]$, $h = 1, ..., p$ and $\epsilon > 0$ as $n \to \infty$.

11.4 Doob-Meyer Composition

Let $X = \{X_0, X_1, X_2, ...\}$ be some general process such that $X_0 = 0$ with respect to history $\{F_n\}$, we define a martingale process $M = \{M_0, M_1, ...\}$ by:

$$M_0 = X_0, \ M_n - M_{n-1} = X_n - E(X_n | F_{n-1})$$

It's clear that $\Delta M_n = M_n - M_{n-1}$ is a martingale difference as the expectation given the past F_{n-1} is zero. This, therefore, can be written as:

$$X_n = E(X_n | F_{n-1}) + \Delta M_n \tag{64}$$

This is the Doob decomposition, where the first term is a function of the past only and the second term ΔN_n is termed *innovations* as it represents new and unexpected information compared to the past.

11.5 Medication used

Table 9 is a clarification of the drugs used in this thesis with their ATC classification and its type.

12 Reference

References

[1] Pazzagli L, Brandt L, Linder M, Myers D, Mavros P, Andersen M, Bahmanyar S. Methods for constructing treatment episodes and impact on exposure-outcome associations. Eur J Clin Pharmacol. 2020 Feb;76(2):267-275. doi: 10.1007/s00228-019-02780-4. Epub 2019 Nov 22. PMID: 31758215.

[2] Wakabayashi R, Hirano T, Laurent T, Kuwatsuru Y, Kuwatsuru R. Impact of "time zero" of Follow-Up Settings in a Comparative Effectiveness Study Using Real-World Data with a Non-user Comparator: Comparison of Six Different Settings. Drugs Real World Outcomes. 2022 Nov 28. doi: 10.1007/s40801-022-00343-1. Epub ahead of print. PMID: 36441486.

[3] Laugesen K, Støvring H, Hallas J, Pottegård A, Jørgensen JOL, Sørensen HT, Petersen I. Prescription duration and treatment episodes in oral glucocorticoid users: application of the parametric waiting time distribution. Clin Epidemiol. 2017 Nov 16; 9:591-600. doi: 10.2147/CLEP.S148671. PMID: 29180903; PMCID: PMC5697451.

ATC	Name	Class
D05BB0	Acitretin	Retinoids for treatment of psoriasis
L04AA06	Mycophenolic Acid	Selective immunosuppressants
L04AA10	Sirolimus	Selective immunosuppressants
L04AA13	Leflunomide	Selective immunosuppressants
L04AA18	Everolimus	Selective immunosuppressants
L04AA21	Efalizumab	Selective immunosuppressants
L04AA23	Natalizumab	Selective immunosuppressants
L04AA24	Abatacept	Selective immunosuppressants
L04AA26	Belimumab	Selective immunosuppressants
L04AA27	Fingolimod	Selective immunosuppressants
L04AA29	Tofacitinib	Selective immunosuppressants
L04AA31	Teriflunomide	Selective immunosuppressants
L04AA32	Apremilast	Selective immunosuppressants
L04AA33	Vedolizumab	Selective immunosuppressants
L04AA37	Baricitinib	Selective immunosuppressants
L04AA40	Cladribine	Selective immunosuppressants
L04AA44	Upadacitinib	Selective immunosuppressants
L04AA45	Filgotinib	Selective immunosuppressants
L04AA50	Ponesimod	Selective immunosuppressants
L04AB01	Etanercept	Tumor necrosis factor alpha (TNF-α) inhibitor
L04AB02	Infliximab	Tumor necrosis factor alpha (TNF-α) inhibitor
L04AB04	Adalimumab	Tumor necrosis factor alpha (TNF-α) inhibitor
L04AB05	Certolizumab pegol	Tumor necrosis factor alpha (TNF-α) inhibitor
L04AB06	Golimumab	Tumor necrosis factor alpha (TNF-α) inhibitor
L04AC01	Daclizumab	Interleukin inhibitors
L04AC03	Anakinra	Interleukin inhibitors
L04AC05	Ustekinumab	Interleukin inhibitors
L04AC07	Tocilizumab	Interleukin inhibitors
L04AC08	Canakinumab	Interleukin inhibitors
L04AC10	Secukinumab	Interleukin inhibitors
L04AC12	Brodalumab	Interleukin inhibitors
L04AC13	Ixekizumab	Interleukin inhibitors
L04AC14	Sarilumab	Interleukin inhibitors
L04AC16	Guselkumab	Interleukin inhibitors
L04AC18	Rinsankizumab	Interleukin inhibitors
L04AC21	Bimekizumab	Interleukin inhibitors
L04AD01	Ciclosporin	Calcineurin inhibitors
L04AD02	Tacrolimus	Calcineurin inhibitors
L04AX01	Azathioprine	Other immunosuppressants
L04AX02	Thalidomide	Other immunosuppressants
L04AX03	Methotrexate	Other immunosuppressants
L04AX04	Lenalidomide	Other immunosuppressants
L04AX05	Pirfenidone	Other immunosuppressants
L04AX06	Pomalidomide	Other immunosuppressants
L04AX07	Dimethyl fumarate	Other immunosuppressants

Table 9: Tabulated list of medication in this thesis

[4] SocialStyrelsen, https://www.socialstyrelsen.se/en/statistics-and-data/registers/national-patient-register/

[5] Gardarsdottir H, Souverein PC, Egberts TC, Heerdink ER. Construction of drug treatment episodes from drug-dispensing histories is influenced by the gap length. J Clin Epidemiol. 2010 Apr;63(4):422-7. doi: 10.1016/j.jclinepi.2009.07.001.

[6] Meaidi M, Støvring H, Rostgaard K, Torp-Pedersen C, Kragholm KH, Andersen M, Sessa M. Pharmacoepidemiological methods for computing the duration of pharmacological prescriptions using secondary data sources. Eur J Clin Pharmacol. 2021 Dec;77(12):1805-1814. doi: 10.1007/s00228-021-03188-9. Epub 2021 Jul 10. PMID: 34247270.

[7] Tanskanen A, Taipale H, Koponen M, Tolppanen AM, Hartikainen S, Ahonen R, Tiihonen J. From prescription drug purchases to drug use periods – a second generation method (PRE2DUP). BMC Med Inform Decis Mak. 2015 Mar 25;15:21. doi: 10.1186/s12911-015-0140-z. PMID: 25890003; PMCID: PMC4382934.

[8] Bharat C, Degenhardt L, Pearson SA, Buizen L, Wilson A, Dobbins T, Gisev N. A data-informed approach using individualised dispensing patterns to estimate medicine exposure periods and dose from pharmaceutical claims data. Pharmacoepidemiol Drug Saf. 2022 Nov 8. doi: 10.1002/pds.5567. Epub ahead of print. PMID: 36345837.

[9] Dr. Johan Reutfors, A Post-Marketing Registry-Based Prospective Cohort Study of Long-Term Safety of Risankizumab in Denmark and Sweden. Centre for Pharmacoepidemiology (CPE), https://www.encepp.eu/encepp/viewResource.htm?id=103640

[10] Odd O. Aalen, Örnulf Borgan, Håkon K. Gjessing. Survival and Event History Analysis. 2008. ISBN 978-0-387-20287-7

[11] David A.Freedman. On The So-Called "Huber Sandvich Estimator" and "Robust Standard Errors". 2006 Sep. Department of Statistics, UC Berkeley

[12] https://www.who.int/tools/atc-ddd-toolkit/methodology