ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis HS 2010

Karin Peter

Marginal Structural Models and Causal Inference

Submission Date: February 28 2011

Adviser: Professor Marloes Maathuis

Abstract

We analyze data of an observational treatment study of HIV patients in Africa, collected by the Institute for Social and Preventive Medicine (ISPM) in Bern. In particular, we focus on patients who received first-line treatment and experienced immunologic failure, where immunologic failure might be an indication that the current treatment is no longer effective. Some of these patients were switched to a second-line treatment, according to the decision of their doctor (i.e. non-randomized). Based on these data, we are interested in estimating the causal effect of the switch to second-line treatment on survival.

The data contain information on the treatment regime and the CD4 counts of the patients, where both of these are time dependent. A main challenge in the analysis is the CD4 count, which indicates how well the immune system is working. The CD4 count may influence future treatment and survival, making it a confounder that one should control for. On the other hand, the CD4 count is likely to be influenced by past treatment, making it an intermediate variable that one should not control for. We address this problem by using marginal structural models. Conceptually, this method weighs each data point by its inverse probability of treatment weight (IPTW), creating data of an unconfounded pseudo-population.

Our results indicate that switching to second-line treatment is beneficial, and slightly more so than an analysis with classical methods would imply.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Description

The institute of Social and Preventive Medicine (ISPM) in Bern, Switzerland supplied us with the data of an observational treatment study of HIV patients in Malawi and Zambia. All patients were treated with so called first line treatment. We consider patients who experienced immunologic failure, meaning that the first line therapy might no longer be effective. Some of these patients were switched to a different treatment (non randomized), called second line treatment. Our aim is to estimate the causal effect of the switch to second line treatment on the survival of the patients.

1.2 Medical Background and Treatment of HIV

We are going to give a brief background about the lapse of an infection with the **human immunodeficiency virus** (HIV) and the standard treatment regime. For any further information about HIV see for example Hartmann (2008) and Deutsche AIDS Gesellschaft e.V. and Osterreichische AIDS Gesellschaft (2010) and the World Health Organization (WHO) homepage (http://www.who.int/hiv/topics/en).

The standard therapy for HIV today is the so called highly active antiretroviral therapy (HAART), where at least three different antiretroviral drugs are combined.

A key point of HIV treatment is the HIV drug resistance. The following definition is from the WHO homepage (http://www.who.int/hiv/topics/drugresistance/en/index.html):

> "The ability of HIV to mutate and reproduce itself in the presence of antiretroviral drugs is called HIV drug resistance. The consequences of drug resistance include treatment failure, increased direct and indirect health costs associated with the need to start more costly second-line treatment for patients, the spread of resistant strains of HIV and the need to develop new anti-HIV drugs. ..."

The viral load is the amount of HIV detectable in the blood. If a HIV patient has non-zero viral load, he is said to experience **virologic failure**. In that case, the patient has to be switched to second line treatment.

One goal of the HAART treatment is to strengthen the immune system, so a unhealthy immune system could be an indicator that the drugs are failing. A typical indicator of the health of the immune system is is the number of CD4 positive cells in the blood: the more cells the healthier the immune system. WHO defined three criteria for so called **immunologic failure**:

- Persistent CD4 count $<$ 100 cells/mm3,
- More than a 50% drop in CD4 count from the peak level,
- A CD4 decrease back to the baseline level or lower.

If any of these three occur, the patient is said to face immunologic failure, and the immune system is no longer able to prevent the body from infections, which could end in the typical clinical picture of AIDS.

In high-income countries the decision to switch treatment is dependent on the viral load and the CD4 count, while patients in middle- and low-income countries do often not have access to viral load testing due to infrastructural and financial reasons. In this case only immunologic failure can be detected. Unfortunately, WHO showed in studies that immunologic and virologic failure does not occur at the same time. The CD4 drop can happen long before actual virologic failure, but can also happen thereafter.

It is important to know, that there is only a limited amount of different lines of treatment, due to a limited amount of antiretroviral pharmaceuticals. Therefore it is not only crucial to switch treatment soon enough, but also not to switch too soon, since at some point there is no more treatment regime left to change to.

A second problem in middle- and low-income countries is that a second line treatment is in general more expensive than first line treatment. So even if immunologic failure is detected, patients do not always get switched to another treatment regime.

1.3 Challenges in the Analysis

The crucial point is that our data is from an observational study, i.e. no double blinded randomized controlled clinical trial and hence we have to deal with confounding. We have to be aware of what covariates to control for and what not, in order to not just receive the correlation but actually the causal effect of the treatment.

Particularly, the problem with this kind of study is that the CD4 count might be a confounder as well as an intermediate variable. This means that it may influence the subsequent treatment, but the CD4 count might itself be influenced by the past treatment. Moreover we assume that it influences the outcome (survival). This circumstance is illustrated for a patient with two follow-up times in Figure 1.1.

On the left panel, the red path demonstrates that the CD4 count is a confounder for treatment at the second follow-up. The red path in the right panel demonstrates that the CD4 cont is an intermediate variable for the treatment at the first follow-up.

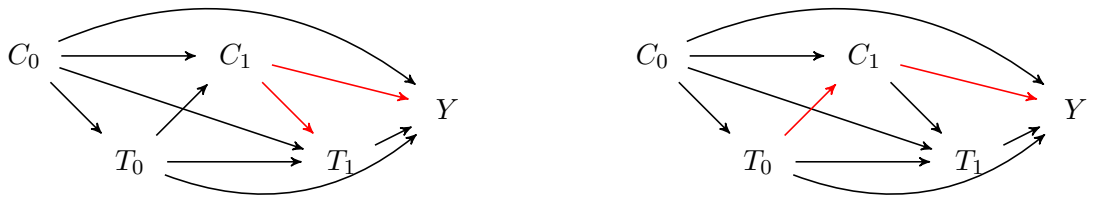

Figure 1.1: Assumed graph of the CD4 count (C) being a confounder (left panel) and a intermediate variable (right panel) for treatment T at the same time.(Y is the indicator variable of death or not death)

1.4 Outline

In the second chapter we are going to address the topic of causality via the do-calculus of Pearl (2000). We give a summary of the most important definitions and equations of Chapter 4 of Pearl (2000). The goal is to understand the formula of the back-door-adjustment and the basic steps to its derivation. This formula is the fundamental equation to address causality.

In the third chapter we explain the Marginal Structural Models (MSM), which was introduced by Robins et al. (2000). The basic idea is to weight the data with the **I**nverse **P**robability of **T**reatment **W**eights (IPTW) to transform the original data into data of a pseudo-population where the treatment is no longer confounded by the time dependent variable. For the special case when all the patients only have one follow-up (so called point-treatment study), we prove that the IPTW method is equivalent to Pearl's back-door-adjustment formula.

In Chapter four, we give a short overview of survival analysis and the use of it in R. In most of the papers that used MSM (like Hernan et al. (2000) for example) the IPTW were estimated using pooled logistic regression. However, Xiao et al. (2010) stated in their paper that using the Cox proportional hazards model leads to better results, hence we focused on this approach. In the appendix the pooled logistic regression is explained briefly and some results of the data analysis with the logistic approach are shown.

Chapter five is about the analysis of our data set. We estimate the weights with the Cox model and accomplish the analysis with the data of the pseudo-population as well as with the original data set and compare the results.

Chapter 2

Causality

2.1 Causality vs. Correlation

An common task in statistics is to estimate the causal effect between two or more random variables. It is very important to make the distinction between pure statistical correlation and causality. In general correlation does not imply a causal effects, while on the other hand causal effects always result in non-zero correlation.

One major problem in the derivation of causal dependence are confounders.

Consider the example illustrated in Figure 2.1:

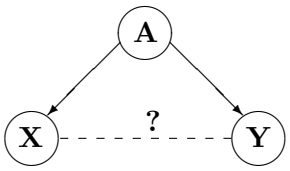

Figure 2.1: Graphical representation of the causal dependence of life expectancy (Y) and the probability of wearing glasses (X)

Let X be the random variable describing if a person has to wear glasses and Y the life expectancy of that person. Then it is likely that we see a statistical dependence between X and Y, but obviously wearing glasses does not influence the life expectancy. In this case age A is a so called confounder, meaning that it is reasonable to believe that the probability of wearing glasses is increasing by age and age influences the life expectancy.

It is therefore important to always think about confounding factors when trying to infer causal effects from observational data. However, if there is dependence between two random variables X and Y it is generally valid to make predictions about the outcome of Y given the value of X even if X and Y have no causal dependence and the correlation is due to confounding.

The classical mathematical tools, only describe the pure statistical dependency structure. Therefore Pearl (2000) introduced a new notation to distinguish between correlation and causality.

Remember that if a random variable X is independent of Y then

$$\mathbf{E}[X|Y] = \mathbf{E}[X]$$

Now assume X influences Y. Then in general

$$\begin{aligned} & \mathbf{E}[X|Y] \neq \mathbf{E}[X] \\ \textit{and} \quad & \mathbf{E}[Y|X] \neq \mathbf{E}[Y] \end{aligned}$$

Hence with this mathematical tool we cannot distinguish if X influences Y or the other way around.

To see this on an example, consider the random variable X representing the kind of disease a person has, Y the kind of drug a doctor subscribes and Z the side effects of a given drug (see Figure 2.2).

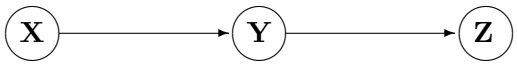

Figure 2.2: Graphical representation of the causal dependence of disease (X), kind of drug (Y) and side effects (Z).

Then:

- $\mathbf{E}[Y|X] \neq \mathbf{E}[Y]$. This is obvious since the kind of disease will influence the choice of the drug a doctor subscribes.
- $\mathbf{E}[X|Y] \neq \mathbf{E}[X]$. Meaning that knowing what drug a person takes tells us something about what disease he might have.

But these two conditional expectations tell us nothing about the fact that the disease affect the choice of drug but the drug don't cause the disease. In order to describe this circumstance, we introduce the following notation:

Definition 2.1.1. *(Pearl (2000), Definition 3.2.1)* ***Causal Effect*** *The causal effect of the random variable X on the random variable Y, denoted as $\mathrm{P}[Y|do(x)]$, is a function from X to the space of probability distributions on Y. For each realization x of X, $\mathrm{P}[Y|do(x)]$ gives the probability of $Y = y$ when X is set to x by some outside intervention.*

Note that this definition can easily be extended to the general case that X and Y are distinct sets of random variables. But for simplicity reasons we are going to restrict ourselves to the case when they are single random variables. An intervention on only one variable is also called *atomic*.

The difference of the above definition to the usual conditional expectation $\mathbf{E}[Y|X = x_0]$ is that in $\mathbf{E}[Y|do(x_0)]$we do not observe that $X = x_0$ but rather intervene to set X to the value x_0 (illustrated in Figure 2.3). Intuitively we can say that $\mathbf{E}[Y|X = x_0]$ is what we can monitor in an observational study, while $\mathbf{E}[Y|do(x_0)]$ can only be observed directly in an experimental study.

Moreover this definition gives rise to the term *counterfactuals* which is due to the fact that in a study we cannot observe the outcome $\mathrm{P}[Y|do(x_0)]$, since the treatment was not

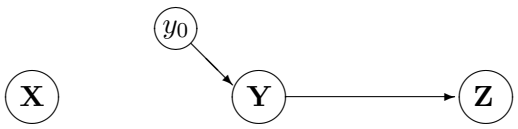

Figure 2.3: Graphical representation of the post-intervention causal dependence of the previous example.

set to x_0 over the whole source population. So $Y|do(x_0)$ is counter to the facts and hence *counterfactual*.

With the example of Figure 2.3 we then have:

- $\mathbf{E}[Y|do(x_0)] = \mathbf{E}[Y|X = x_0]$ Because giving a person a specific disease will lead to the same choice of drug as if we would just observe somebody already having the disease.
- $\mathbf{E}[X|do(y_0)] = \mathbf{E}[X]$ Implying that forcing a person to take a specific drug will still not give us any more information about what disease he has (if any).

We will refer to $\mathbf{E}[Y|X]$ as the pre-intervention expected value and to $\mathbf{E}[Y|do(x_0)]$ as the post-intervention expected value.

Especially in drug treatment studies the value of interest is mainly the post-intervention expected value, but in observational studies only the pre-intervention expected values can be observed. Hence we need to derive formulas to receive the desired values out of the observed ones.

2.2 Basic Notations, Definitions and Formulas

In the following we will recall some basic notations, definitions and formulas based on Pearl (2000).

Recall that we are interested in the influence of one variable on another. It is important to know all the factors that influence or are influenced by one of the variables of interest. Pearl (2000) therefore introduce causal diagrams as a useful tool to determine causal inference.

Definition 2.2.1. *A causal diagram is a graph where all the variables are represented by vertices and there is an arrow from one variable X to another Y if X has an effect on Y.*

Note that the most substantial assumptions of a causal diagram are not founded in the arrows , but rather in the missing arrows. I.e. an arrow from X to Y implies that there might be an causal effect between the two, but a missing arrow tells us that it is sure that there is no such interaction.

We do not just need these graphs to visualize a causal interaction, as we will see, it is rather a crucial mathematical tool that we are going to use to determine the post-intervention distribution of a random variable.

We are only going to deal with models that can be represented by directed acyclic graphs (DAG).

An useful equation for the following derivations is the chain rule:

$$\mathrm{P}\left[X_1, \ldots, X_n\right] = \prod_{i=1}^{n} \mathrm{P}\left[X_i | X_1, \ldots X_{i-1}\right]$$

where the set $\{X_0 = x_0\}$ is the empty set.

The chain rule follows directly from the definition of conditional probability $\mathrm{P}\left[A|B\right] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]}$.

The following definition gives us another useful tool to simplify the chain rule.

Definition 2.2.2. *(Pearl (2000), Definition 1.1.2) Let* $\mathrm{P}\left[\cdot\right]$ *be a joint probability function and let* X*,* Y*, and* Z *be any variables. The variables* X *and* Y *are said to be conditionally independent given* Z *if*

$$\mathrm{P}\left[X|Y, Z\right] = \mathrm{P}\left[X|Z\right], \ \textit{whenever} \ \mathrm{P}\left[Y, Z\right] > 0$$

In the example in Figure 2.2 we have that Z and X are conditionally independent given Y, because the kind of disease does not give us any more information about the side effects of the drug if we know the drug the patient takes.

For a graphical interpretation see also Figure 2.4.

(a) $\mathrm{P}\left[Z|X, Y\right] = \mathrm{P}\left[Z|Y\right]$ (b) $\mathrm{P}\left[Z|X, Y\right] \neq \mathrm{P}\left[Z|Y\right]$

Figure 2.4: In graph (a) Z is conditionally independent of X given Y, while in graph (b) they are not conditionally independent.

With this definition we can simplify each term on the right of the chain rule by conditioning only on the subset $PA_i \subseteq \{X_1, \ldots, X_{i-1}\}$ of variables where PA_i is such that $\mathrm{P}\left[X_i|X_1, \ldots, X_{i-1}\right] = \mathrm{P}\left[X_i|PA_i\right]$, meaning that X_i and $\{X_1, \ldots, X_{i-1}\} \setminus PA_i$ are conditionally independent given PA_i.
We then get the new version of the chain rule:

$$\mathrm{P}\left[X_i|X_1, \ldots X_{i-1}\right] = \prod_{i=1}^{n} \mathrm{P}\left[X_i|PA_i\right] \tag{2.2.0.1}$$

The notation PA_i comes from the fact that all the variables in PA_i are actually those variables of $\{X_1, \ldots, X_i\}$ that have an arrow pointing towards X_i ,i.e. the parents graphically speaking. Let pa_i be all the parents of X_i, then it turns out that we can replace PA_i by pa_i. This is not obvious because PA_i of X_i are defined to be a subset of $\{X_1, \ldots, X_{i-1}\}$, whereas the variables of pa_i are in general not all in $\{X_1, \ldots, X_{i-1}\}$. To get that $PA_i = pa_i$ we need an appropriate ordering of the variables, such that $pa_i \subseteq \{X_i, \ldots, X_{i-1}\}$ for all $i \in \{1, \ldots, n\}$. One can show that this is indeed feasible in graphs without directed cycles.

From now on we will assume we have such an ordering. We denote the parents of the random variable X_i as PA_i and the realizations of PA_i as pa_i.

Corresponding to a DAG we can define a structural equation model where each function correspond to the child-parents relationship in the graph:

$$x_i \leftarrow f_{x_i}(pa_i, \epsilon_i),$$

where pa_i are the parents of x_i and ϵ_i are arbitrary random disturbances. Then the marginal probability distribution $\mathrm{P}[X_i|pa_i]$ only depends on the function f_{X_i} and not on the other functions f_{X_j}, whenever $j \neq i$.

The ϵ_i are assumed to be independent. If there is any reason to believe that the ϵ_i are not independent, they have to be taken into the model by adding an unmeasured variable U.

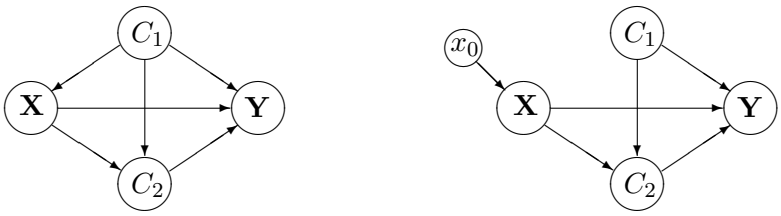

Figure 2.5: Graphs corresponding to the non-parametric model Figure 2.2.0.2 and Figure 2.2.0.3.

We are first going to intuitively derive the so called truncated factorization formula using the example of Figure 2.5. The left graph of Figure 2.5 corresponds to:

$$X \leftarrow f_X(C_1, \epsilon_X), \quad C_1 \leftarrow f_{C_1}(\epsilon_{C_1}), \quad C_2 \leftarrow f_{C_2}(X, C_1, \epsilon_{C_2}), \quad Y \leftarrow f_Y(X, C_1, C_2, \epsilon_Y). \tag{2.2.0.2}$$

Whereas the post-intervention model after an outside intervention at X is given in the right graph of Figure 2.5. It gives rise to the equations:

$$X \leftarrow x_0, \quad C_1 \leftarrow f_{C_1}(\epsilon_{C_1}), \quad C_2 \leftarrow f_{C_2}(x_o, C_1, \epsilon_{C_2}), \quad Y \leftarrow f_Y(x_o, C_1, C_2, \epsilon_Y). \tag{2.2.0.3}$$

Assume that ϵ_X,ϵ_Y, ϵ_{C_1} and ϵ_{C_2} are mutually independent. By (2.2.0.1) we can now calculate the post-intervention distribution of Figure 2.5.

Note first:

$$\begin{aligned}\mathrm{P}[X, C_1, C_2, Y] &= \mathrm{P}[X|PA_X] \cdot \mathrm{P}[C_1|PA_{C_1}] \cdot \mathrm{P}[C_2|PA_{C_2}] \cdot \mathrm{P}[Y|PA_Y] \\ &= \mathrm{P}[X|C_1] \cdot \mathrm{P}[C_1] \cdot \mathrm{P}[C_2|X, C_1] \cdot \mathrm{P}[Y|X, C_1, C_2]\,.\end{aligned}$$

Intuitively the post-intervention distribution is then:

$$\mathrm{P}[X, C_1, C_2, Y|do(x_0)] = \begin{cases} 0 & \text{if } X \neq x_0 \\ \mathrm{P}[C_1] \cdot \mathrm{P}[C_2|x_0, C_1] \cdot \mathrm{P}[Y|x_0, C_1, C_2] & \text{if } X = x_0 \end{cases}$$

because if we intervene to set $X = x_0$ then we observe $X = x_0$ with probability one, i.e.:

$$\mathrm{P}[X = x] = \begin{cases} 0 & \text{if } x \neq x_0 \\ 1 & \text{if } x = x_0. \end{cases}$$

This generalizes to the Truncated factorization formula:

$$\mathrm{P}[X_1, \ldots, X_n|do(x_i)] = \begin{cases} 0 & \text{if } X_i \neq x_i \\ \prod_{j \neq i} \mathrm{P}[X_j|PA_j] & \text{if } X_i = x_i \end{cases}, \tag{2.2.0.4}$$

which can be rewritten as:

$$\begin{aligned}
\mathrm{P}\left[X_1,\ldots,X_n|do(x_i)\right] &= \prod_{j\neq i}\mathrm{P}\left[X_j|PA_j\right]\cdot\frac{\mathrm{P}\left[X_i|PA_i\right]}{\mathrm{P}\left[X_i|PA_i\right]}\\
&= \prod_{j=1}^{n}\mathrm{P}\left[X_j|PA_j\right]\frac{1}{\mathrm{P}\left[X_i|PA_i\right]}\\
&= \frac{\mathrm{P}\left[X_1,\ldots,X_n\right]}{\mathrm{P}\left[X_i|PA_i\right]}\text{ , for } X_i = x_i.
\end{aligned}$$

This formula can be interpreted as follows: $\mathrm{P}\left[X_i = x_i|pa_i\right]$ is the probability of observing X_i to be x_i given all the values of its predecessors. So by dividing by $\mathrm{P}\left[X_i = x_i|pa_i\right]$ the probability of the joint distribution $\mathrm{P}\left[x_1,\ldots,x_n\right]$ is increased by a factor $\mathrm{P}\left[X_i = x_i|pa_i\right]^{-1}$. I.e. we calculate the joint probability and then adjust for the fact that we forced X_i to be x_i rather than observed it.

Further we can write (2.2.0.4) the following way by using $\mathrm{P}\left[x_i|pa_i\right] = \frac{\mathrm{P}[x_i,pa_i]}{\mathrm{P}[pa_i]}$:

$$\begin{aligned}
\mathrm{P}\left[X_1,\ldots,X_n|do(x_i)\right] &= \frac{\mathrm{P}\left[X_1,\ldots,X_n\right]}{\mathrm{P}\left[X_i|PA_i\right]}\\
&= \frac{\mathrm{P}\left[X_1,\ldots,X_n\right]}{\mathrm{P}\left[X_i,PA_i\right]}\cdot\mathrm{P}\left[PA_i\right]\\
&= \mathrm{P}\left[X_1,\ldots,X_n|X_i,PA_i\right]\cdot\mathrm{P}\left[PA_i\right].
\end{aligned}\tag{2.2.0.5}$$

For discrete variables we then get the formula:

Theorem 2.2.3. *(Pearl (2000), Theorem 3.2.2)* ***Adjustment for direct causes*** *Let PA_i denote the set of direct causes on the variable X_i, and Y any variable disjoint of $\{X_i \cup PA_i\}$. The effect of the intervention $do(X_i = x_i')$ on Y is given by*

$$\mathrm{P}\left[Y = y|do(x_i')\right] = \sum_{pa_i}\mathrm{P}\left[Y = y|X_i = x_i', PA_i = pa_i\right]\cdot\mathrm{P}\left[PA_i = pa_i\right].$$

Where $\mathrm{P}\left[Y = y|X_i = x_i', PA_i = pa_i\right]$ and $\mathrm{P}\left[PA_i = pa_i\right]$ represent pre-intervention probabilities.

Proof. Let $Y \in \{X_1,\ldots,X_n\}\setminus X_i$. Then the formula is derived by summing (2.2.0.5) over all realizations of the variables $\{X_1,\ldots,X_n\}\setminus\{X_i,Y\}$ and the fact that $\mathrm{P}\left[X\right] = \sum_y \mathrm{P}\left[X,Y=y\right]$ for discrete variables:

$$\sum_{\{x_1,\ldots,x_n\}\setminus\{x_i,y\}}\mathrm{P}\left[X_1,\ldots,X_n|do(x_i)\right] = \mathrm{P}\left[Y,X_i|do(x_i)\right] = \mathrm{P}\left[Y|do(x_i)\right]$$

and

$$\begin{aligned}
&\sum_{\{x_1,\ldots,x_n\}\setminus\{x_i,y\}} \mathrm{P}\left[X_1,\ldots,X_n|x_i,PA_i\right]\cdot\mathrm{P}\left[PA_i\right]\\
&=\sum_{pa_i}\sum_{\{x_1,\ldots,x_n\}\setminus\{x_i,y,pa_i\}} \mathrm{P}\left[PA_i\right]\cdot\mathrm{P}\left[X_1,\ldots,X_n|x_i,PA_i\right]\\
&=\sum_{pa_i}\mathrm{P}\left[PA_i\right]\sum_{\{x_1,\ldots,x_n\}\setminus\{x_i,y,pa_i\}} \mathrm{P}\left[X_1,\ldots,X_n|x_i,PA_i\right]\\
&=\sum_{pa_i}\mathrm{P}\left[PA_i\right]\cdot\mathrm{P}\left[Y,X_i,PA_i|x_i,PA_i\right]\\
&=\sum_{pa_i}\mathrm{P}\left[Y|x_i,PA_i\right]\cdot\mathrm{P}\left[PA_i\right].
\end{aligned}$$

□

This formula is called *adjusted for direct causes* and can be interpreted as conditioning on the direct causes and then averaging over all realization of PA_i. Note that this formula is useful in practice only if there are no unmeasured variables affecting X_i. In case of unmeasured variables in the set PA_i the values $\mathrm{P}\left[Y=y|X_i,PA_i\right]$ and $\mathrm{P}\left[PA_i\right]$ are not known.

Also note that the assumption $Y\notin\{X_i\cup PA_i\}$ is essential and it means that the outcome is not allowed to have a influence on any of the parents of X_i.

2.3 Controlling for Confounders

Recall that in Figure 2.1 an analysis of the variable X (wearing glasses Yes/No) on Y (life expectancy) would reveal a positive correlation between the two even though they do not have a causal effect on each other. Making an analysis of X and age A on Y will give us the true causal coefficient for X.

This can be understood the following way: If we only regress on X, then X will carry also the amount of influence that A has on Y, not only its own amount of influence. But if we take both variables into account, then the influence of A on Y is represented by the parameter of A itself. This approach is called *controlling for a confounder*.

Observe that the formula of Theorem 2.2.3 does nothing else than controlling for all the confounders. The problem is that in practice we often do not know which variables we have to control for.

First of all it is important to notice that we will not avoid the problem of choosing an appropriate set of variables to control for if we would just control for all the variables. We only want to control for the confounders, but not for the intermediates. If we would control for an intermediate the same problem described above would occur, i.e. an amount of the influence of X on Y would be displayed in the parameter of the intermediate.

Example 2.3.1. To see the impact of spuriously controlling for intermediates, consider the causal model represented in Figure 2.6 with the following linear structural equations:

$$\begin{aligned}
C_1&=3\cdot X,\\
C_2&=5\cdot C_1+8,\\
Y&=2\cdot X-2\cdot C_1+4\cdot C_2
\end{aligned}$$

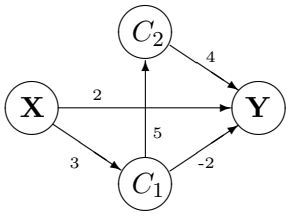

Figure 2.6: Graphical representation of Example 2.3.1.

If we want to determine the causal effect (total effect) of X on Y, then we should get 56 for the slope and 32 the intercept, as we can see by the calculation:

$$\begin{aligned} Y &= 2 \cdot X - 2 \cdot C_1 + 4 \cdot C_2 \\ &= 2 \cdot X - 2 \cdot (3 \cdot X) + 4 \cdot (5 \cdot C_1 + 8) \\ &= 2 \cdot X - 6 \cdot X + 4 \cdot (15 \cdot X + 8) \\ &= 56 \cdot X + 32. \end{aligned}$$

But if we would take both C_1 and C_2 into the model, we would just get 2 as the slope.

Note that the causal effect is the total effect that X has on Y. If we include C_1 and C_2 in the model, then they both have an effect on Y, but whatever effect they have on Y is only due to the effect that X has on them.

As a second problem on the task of controlling is that we have to be aware that there might be unmeasured confounders and controlling only on the measured confounders will lead to biased estimates.

In the previous section we already stated a formula where we just control for all the graphical parents of X. In fact this might not be a minimal set of variables we need to control for, so especially if there are direct unmeasured causes we would want to know if there is some other set of variables which is sufficient for controlling.

The so called *Back-Door Adjustment* specifies sufficient conditions for a set of variables to be controlled on. We first need two more definitions:

Definition 2.3.2. *(Pearl (2000), Definition 1.2.3)**Block a path** The path p in the graph G is said to be blocked by the set of vertices $Z \subset G$ if and only if*

- *p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z.*
  or
- *p contains an inverted fork (or collider) $i \to m \leftarrow J$ such that the middle node m is not in Z and such that no descendant of m is in Z.*

For an example see Figure 2.7.

Definition 2.3.3. *(Pearl (2000),Definition 3.3.1) **Back-Door Criterion** A set of variables Z satisfy the back-door criterion relative to the ordered pair (X_i, X_j) in a DAG if:*

- *no node in Z is a descendant of X_i; and*

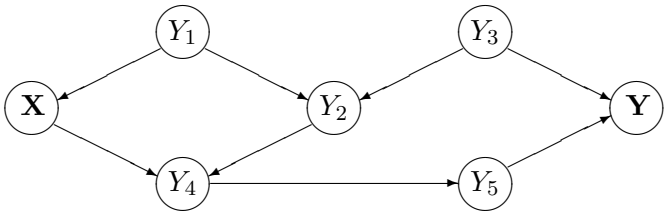

Figure 2.7: Blocking paths:

- The path $X - Y_1 - Y_2 - Y_3 - Y$ is blocked if either $Y_2, Y_4 \notin Z$ or at least one of $Y_1, Y_3 \in Z$.
- The path $X - Y_1 - Y_2 - Y_4 - Y_5 - Y$ is blocked if at least one of Y_1, Y_2, Y_4, Y_5 is in Z.
- The path $X - Y_4 - Y_2 - Y_3 - Y$ is blocked if Y_4 and $Y_5 \notin Z$ or if at least one of $Y_2, Y_3 \in Z$.
- The path $X - Y_4 - Y_5 - Y$ is blocked if at least one of $Y_4, Y_5 \in Z$.

- *Z blocks every path between X_i and X_j that contains an arrow into X_i.*

In Figure 2.7 the two paths $X - Y_1 - Y_2 - Y_3 - Y$ and $X - Y_1 - Y_2 - Y_4 - Y_5 - Y$ have to be blocked by Z in order to satisfy the back door criterion relative to (X, Y). By definition Y_4 cannot be in Z. So if also Y_2 is not in Z, then the first path is already blocked. To block the second path we could add Y_1 or Y_5 to Z. Hence $\{Y_1\}$ and $\{Y_5\}$ satisfy the back-door criterion relative to (X, Y).

Finally we have the following theorem:

Theorem 2.3.4. *(Pearl (2000), Theorem 3.3.2)* ***Back-Door Adjustment*** *If a set of random variables Z satisfies the back-door criterion relative to (X, Y) and no variables of Z are unmeasured, then the causal effect of X on Z is identifiable and is given by the formula*

$$\mathrm{P}\left[Y = y | do(x_0)\right] = \sum_z \mathrm{P}\left[Y = y | X = x, Z = z\right] \mathrm{P}\left[Z = z\right] \tag{2.3.0.6}$$

Chapter 3

Marginal Structural Models

Robins et al. (2000) introduced Marginal structural models as an alternative to the formula of the back-door-adjustment to obtain causal inference from observational data. Among others, this approach has been used by Hernan et al. (2000) and Sterne et al. (2005). Xiao et al. (2010) did a simulation study to compare conventional methods with MSM.

Basically this method is about weighting the data by the inverse-probability-of-treatment weights (IPTW) to construct a pseudo-population which is no longer confounded. I.e. we simulate data of an experimental study out of the data of an observational study. To estimate the causal effect of the treatment we can then apply any model to the data of the pseudo-population without caring for confounders.

3.1 Time Dependent Treatment Studies

For a time dependent treatment study we assume a graph of the form represented in Figure 3.1. Note that in this graph we only have two time-point, while in reality there can be an arbitrary amount of time-points. The treatment T changes over time and there are several indicators C (also changing over time) influencing the decision of what treatment to choose and also being indicators for the outcome Y. Moreover the treatment itself influences the indicators. U are unmeasured variables. Consider for example T_i to be the use of a beta-blocker at time i, C_i the blood pressure at time i and Y the occurrence of a heart attack at the end of time n.

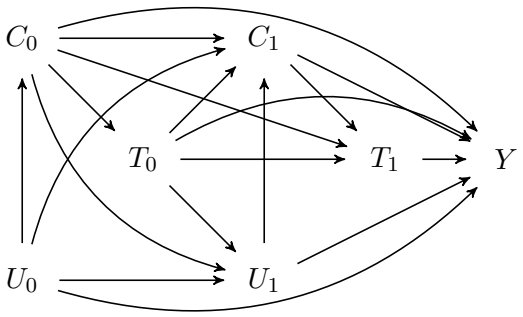

Figure 3.1: Assumed graph for the use of Marginal Structural Models.

Definition 3.1.1. *A **Time-dependent confounder** is a time dependent covariate which is a predictor of the outcome as well as a predictor of subsequent treatment.*

If in a study there is a time dependent confounder which is also influenced by past treatment history (as in Figure 3.1), then standard approaches are biased as already mentioned in the previous chapter.

More specifically, if we are interested in the effect of treatment regime $do(\overline{T} = (t_1, \ldots, t_n))$ on the outcome Y, we would have to control for all the confounders C_i, but since they are influenced by past treatment, they are also intermediates, and we would not want to control for them. This leads to a dilemma and we cannot just do a straightforward analysis. Pearl (2000) (Chapter 4) does have another solution to this problem, but we are going to focus on the Marginal Structural Model approach.

Remember that in graphs like Figure 3.1 the important assumptions lie in the missing arrows. So there are three important assumptions we make in this graph in order for MSM to work:

- There is no arrow from unmeasured variables to the treatment, meaning that we assume there are no unmeasured confounders for the treatment,
- There are no arrows from future treatment and/or confounder to past treatment and/or confounder,
- There is no arrow from the outcome Y to the treatment.

3.2 Point Treatment Study

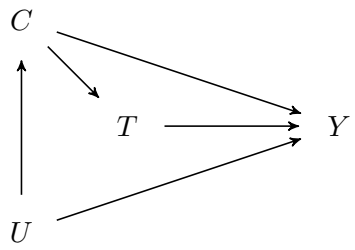

Figure 3.2: Point treatment study where U are unmeasured variables, C a confounder for the treatment, T the treatment and Y the outcome.

Let us first cover the case where we only have a one-time treatment decision as represented in Figure 3.2. U stands for the unmeasured covariates, and the missing arrows from U to T implies that there are no unmeasured confounders for T. In reality it is in general not at all reasonable to make this assumption, but we need to make it in order to be able to do the calculations.

As already mentioned, we will weight the data in order to construct a pseudo-population of data that would have been collected out of an unconfounded or experimental study. More precisely each person contributes $w_i = \mathrm{P}\left[T = t|C = c\right]^{-1}$ copies of itself to the population, i.e. the inverse of the probability of receiving its prescribed treatment given its covariate values.

Let us first look at an example with a discrete treatment variable and a dichotomous outcome. In practical applications this would mainly refer to the outcome Y representing if the person is dead or not, or if some illness is cured or not.

The value of interest is then the causal effect of the treatment $T = t_0$ on outcome Y, i.e the probability of surviving when treated with treatment t_0.

Some other values of interest are:

- Causal risk difference: $\mathrm{P}[Y = 1|do(t_1)] - \mathrm{P}[Y = 1|do(t_0)]$
- Causal risk ratio: $\mathrm{P}[Y = 1|do(t_1)] \,/\, \mathrm{P}[Y = 1|do(t_0)]$
- Causal odds ratio: $\mathrm{P}[Y = 1|do(t_1)]\cdot\mathrm{P}[Y = 0|do(t_0)] \,/\, \mathrm{P}[Y = 1|do(t_0)]\cdot\mathrm{P}[Y = 0|do(t_1)]$

These values basically are different measures of the increase of life expectancy if we were to use treatment t_1 rather then treatment t_0 over the entire source population.
Let us first look at an example before proving that this weighting actually does the trick.

Example 3.2.1. Consider the example where $T = 1$ would imply treatment with a beta-blocker ($T = 0$ implying no drug used), C the blood pressure and $Y = 1$ whenever a heart attack occurs. A treatment regime could be that the doctors are more likely to start treatment when the blood pressure pass over a certain critical value p_{crit}.

In this case, we are going to observe only few people with high blood pressure and no treatment. There could actually be significantly less than if we would have given treatment randomly. Hence we have to adjust for the confounders by adding data of 'faked' people with high blood pressure and no treatment such that there are enough of these people to consider the treatment to be randomly given.

Hypothetical data from an observational study with eight patients is given in Table 3.1.

	$C \geq p_{crit}$		$C < p_{crit}$	
	$T = 1$	$T = 0$	$T = 1$	$T = 0$
$Y = 0$	3	0	1	2
$Y = 1$	1	1	0	0
Total	4	1	1	2

Table 3.1: Hypothetical data for Example 3.2.1.

Note that:

$$\mathrm{P}[Y = 0|T = 1] = \frac{\mathrm{P}[Y = 0, T = 1]}{\mathrm{P}[T = 1]} = \frac{4}{8} : \frac{5}{8} = \frac{4}{5} = 0.80, \text{ and}$$
$$\mathrm{P}[Y = 0|T = 0] = \frac{\mathrm{P}[Y = 0, T = 0]}{\mathrm{P}[T = 0]} = \frac{2}{8} : \frac{3}{8} = \frac{2}{3} = 0.67$$

Then the crude risk difference $\mathrm{P}[Y = 0|T = 1] - \mathrm{P}[Y = 0|T = 0]$ is $\frac{2}{15} = 0.13$, which would imply that the probability of having a heart attack would only decrease minimally with the use of beta-blocker.

In contrast, the formula of Theorem 2.2.3 yields:

$$\begin{aligned}
\mathrm{P}\left[Y=0|do(T=1)\right] &= \mathrm{P}\left[Y=0|T=1, C\geq p_{crit}\right]\cdot\mathrm{P}\left[C\geq p_{crit}\right] \\
&\quad + \mathrm{P}\left[Y=0|T=1, C< p_{crit}\right]\cdot\mathrm{P}\left[C< p_{crit}\right] \\
&= \frac{3}{4}\cdot\frac{5}{8}+\frac{1}{1}\cdot\frac{3}{8}=\frac{27}{32}=0.84375, \text{ and} \\
\mathrm{P}\left[Y=0|do(T=0)\right] &= \mathrm{P}\left[Y=0|T=0, C\geq p_{crit}\right]\cdot\mathrm{P}\left[C\geq p_{crit}\right] \\
&\quad + \mathrm{P}\left[Y=0|T=0, C< p_{crit}\right]\cdot\mathrm{P}\left[C< p_{crit}\right] \\
&= 0\cdot\frac{5}{8}+1\cdot\frac{3}{8}=\frac{3}{8}=0.375
\end{aligned}$$

And the causal risk difference $\mathrm{P}\left[Y=0|do(T=1)\right]-\mathrm{P}\left[Y=0|do(T=0\right]$ is 0.46875, which indicates a much larger effect.

As already mentioned, this effect comes from the fact that there are not the same amount of patients in the two treatment groups $T=1$ and $T=0$ within each confounder level. For example, there are four patients with a high blood pressure receiving treatment but only one not receiving treatment, while in an unconfounded experimental study, there would be about the same amount of patients in each group.

Let us check if we get the same value (0.46875) by weighting with the IPTW.
The weight for the group with $C\geq p_{crit}$ and $T=1$ (i.e. the weight for the first column in the table) is:

$$\mathrm{P}\left[T=1|C\geq p_{crit}\right]^{-1}=\left(\frac{\mathrm{P}\left[T=1, C\geq p_{crit}\right]}{\mathrm{P}\left[C\geq p_{crit}\right]}\right)^{-1}=\left(\frac{4}{8}:\frac{5}{8}\right)^{-1}=\left(\frac{4}{5}\right)^{-1}=\frac{5}{4}$$

Similarly we can calculate all the other weights:

$$\mathrm{P}\left[T=0|C\geq p_{crit}\right]^{-1}=\left(\frac{\mathrm{P}\left[T=0, C\geq p_{crit}\right]}{\mathrm{P}\left[C\geq p_{crit}\right]}\right)^{-1}=\left(\frac{1}{8}:\frac{5}{8}\right)^{-1}=\left(\frac{1}{5}\right)^{-1}=5$$

$$\mathrm{P}\left[T=1|C< p_{crit}\right]^{-1}=\left(\frac{\mathrm{P}\left[T=1, C< p_{crit}\right]}{\mathrm{P}\left[C< p_{crit}\right]}\right)^{-1}=\left(\frac{1}{8}:\frac{3}{8}\right)^{-1}=\left(\frac{1}{3}\right)^{-1}=3$$

$$\mathrm{P}\left[T=0|C< p_{crit}\right]^{-1}=\left(\frac{\mathrm{P}\left[T=0, C< p_{crit}\right]}{\mathrm{P}\left[C< p_{crit}\right]}\right)^{-1}=\left(\frac{2}{8}:\frac{3}{8}\right)^{-1}=\left(\frac{2}{3}\right)^{-1}=\frac{3}{2}$$

	$C\geq p_{crit}$		$C< p_{crit}$	
	$T=1$	$T=0$	$T=1$	$T=0$
$Y=0$	3.75	0	3	3
$Y=1$	1.25	5	0	0
Total	5	5	3	3
Weights	$\frac{5}{4}$	5	3	$\frac{3}{2}$

Table 3.2: Data of the pseudo-population.

The data of the pseudo-population and all the weights are are listed in Table 3.2. From the original data to the data of the pseudo-population we just have to multiply all the numbers in Table 3.1 with the corresponding weights.

In the pseudo-population we have:

$$\mathrm{P}\left[Y=0|T=1\right]=\frac{\mathrm{P}\left[Y=0,T=1\right]}{\mathrm{P}\left[T=1\right]}=\frac{6.75}{16}:\frac{8}{16}=\frac{6.75}{8}=0.84375$$
$$\mathrm{P}\left[Y=0|T=0\right]=\frac{\mathrm{P}\left[Y=0,T=0\right]}{\mathrm{P}\left[T=0\right]}=\frac{3}{16}:\frac{8}{16}=\frac{3}{8}=0.375,$$

which results in the risk difference of 0.46875 as claimed.

We see in the table that the study is no longer confounded by the blood pressure. In both blood pressure groups 50% of the patients receiving treatment and 50% do not. Hence the crude effect of treatment in the pseudo-population is equal to the causal effect in the real population.

We will now prove for a point treatment study, that the IPTW-method does give us the causal effect by showing that it is equivalent to pearl's do-calculus.

Theorem 3.2.2. *If the data of a study which is given by the causal graph of Figure 3.2 (or a subgraph of it) is weighted by the IPT Weights* $w_i=\mathrm{P}\left[T=t_i|C=c_i\right]^{-1}$*, where* t_i *and* c_i *are the values of* T *and* C *corresponding to the* $i-th$ *person, and the data of the weighted population is denote by* T^p*,* C^p *and* Y^p*, then*

$$\mathrm{P}\left[Y=y|do(T=t)\right]=\mathrm{P}\left[Y^p=y|T^p=t\right],\ \textit{for } Y \textit{ discrete.}$$

Proof. Note that $\mathrm{P}\left[Y^P=y,T^P=t,C^P=c\right]$ is the relative frequency of the event $Y^P=y,T^P=t,C^P=c$, so $\mathrm{P}\left[Y^P=y,T^P=t,C^P=c\right]=\frac{\mathrm{P}[Y=y,T=t,C=c]}{\mathrm{P}[T=t|C=c]}$. Then:

$$\begin{aligned}
\mathrm{P}\left[Y^p=y|T^p=t\right]&=\frac{\mathrm{P}\left[Y^p=y,T^p=t\right]}{\mathrm{P}\left[T^p=t\right]}\\
&=\frac{\sum_k \mathrm{P}\left[Y^p=y,T^p=t,C^p=c_k\right]}{\sum_k \mathrm{P}\left[T^p=t,C^p=c_k\right]}\\
&=\frac{\sum_k \mathrm{P}\left[Y=y,T=t,C=c_k\right]\cdot w_k}{\sum_k \mathrm{P}\left[T=t,C=c_k\right]\cdot w_k}\\
&=\frac{\sum_k \mathrm{P}\left[Y=y|T=t,C=c_k\right]\cdot \mathrm{P}\left[T=t,C=c_k\right]\cdot \frac{\mathrm{P}[C=c_k]}{\mathrm{P}[T=t|C=c_k]}}{\sum_k \mathrm{P}\left[T=t,C=c_k\right]\cdot \frac{\mathrm{P}[C=c_k]}{\mathrm{P}[T=t,C=c_k]}}\\
&=\frac{\sum_k \mathrm{P}\left[Y=y|T=t,C=c_k\right]\cdot \mathrm{P}\left[C=C_k\right]}{\sum_k \mathrm{P}\left[C=c_k\right]}\\
&=\sum_k \mathrm{P}\left[Y=y|T=t,C=c_k\right]\cdot \mathrm{P}\left[C=c_k\right].
\end{aligned}$$

Which is exactly the formula of Theorem 2.2.3.

Note that it is important that there is no arrow from Y to C, as in Theorem 2.2.3 the outcome is not allowed to have a causal influence on the parents of T. □

Notes on the weighting:

i.) There has to be at least one observation in each group to be able to calculate $\mathrm{P}\left[T=t|C=c\right]$ directly, otherwise $\mathrm{P}\left[T=t|C=c\right]$ might be zero for some values of T and C, therefore the weights would be undefined.

ii.) These weights can become extremely large, meaning that a small group of individuals contribute a large amount of data. The estimates will then have large variance.

If we have the first problem or also if we have continuous confounders C, we are not able to calculate the weights directly as the ratio of amount of beneficial events and total amount of events. In this case we will have to specify a model for $T|C = c$. Two important examples for outcomes in practice are dichotomous outcomes and survival times. In the first case we could choose a logistic model, while for survival times Cox proportional hazard models could be used.

The second problem can be improved by using stabilized weights $sw_i = \frac{\mathrm{P}[T=t]}{\mathrm{P}[T=t|C=c]}$ instead of w_i. The values of sw_i oscillate in a more narrower range around one. We can directly see in the proof of Theorem 3.2.2 that also with sw_i we get the desired property. The only difference is that in line three of the calculation, the nominator of sw_i would cancel out since it is not dependent on the summation index.

Remark: Xiao et al. (2010) suggested another definition for more stabilized weights, so called normalized stabilized weights, and showed in their simulation study that the results were slightly better.

3.3 Time Dependent Studies

The previous section can be generalized to time dependent models as in Figure 3.1 where we possibly have more than two time-points.
Robins et al. (2000) showed that $\mathrm{P}\left[Y = y|do(t_1, \ldots, t_n)\right]$ can be calculated as the crude effect $\mathrm{P}\left[Y = y|\overline{T}_n = \overline{t}_n\right]$ in the pseudo population derived from the data by weighting it with the weights

$$w = \prod_{k=0}^{K} \mathrm{P}\left[T_k = t_k|\overline{T}_{k-1} = \overline{t}_{k-1}, \overline{C}_k = \overline{c}_k\right]^{-1}$$

where $\overline{T}_k = (T_1, \ldots, T_k)$, and $\overline{C}_k = (C_1, \ldots, C_k)$.
And the stabilized weights are:

$$sw = \prod_{k=0}^{K} \frac{\mathrm{P}\left[T_k = t_k|\overline{T}_{k-1} = \overline{t}_{k-1}\right]}{\mathrm{P}\left[T_k = t_k|\overline{T}_{k-1} = \overline{t}_{k-1}, \overline{C}_k = \overline{c}_k\right]}.$$

3.4 Multiple Treatments

It is possible that there are multiple treatments involved, that might or might not be dependent on each other. An example of that could be that a patient with a heart disease has the possibility of been treated with drugs, or with an operation, or even both. In Figure 3.3 T stands for the treatment with drugs and T' for the operation. The two sided arrow between T and T' implies that they could be correlated in some way. For simplicity we omitted some of the arrows in this graph. But be aware, that whenever there is a directed path between two knots, there should also be direct arrow.

In the presence of multiple treatments we are actually interested in the survival probability given that the first treatment regime is set to $\overline{t}$ and the second treatment regime is set to $\overline{t'}$, i.e. $\mathrm{P}\left[Y = 0|do(\overline{T} = \overline{t}), do(\overline{T'} = \overline{t'})\right]$.

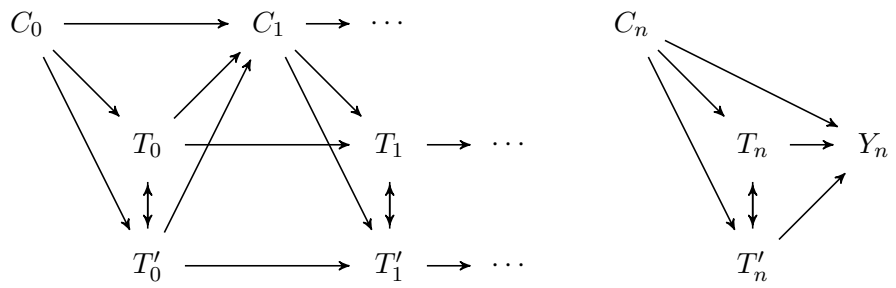

Figure 3.3: Time dependent treatment study with multiple treatments.

As we can calculate with the definition of conditional probability, the weight for patient i becomes:

$$\begin{aligned} w_i^{-1} &= \prod_{k=0}^{n} \mathrm{P}\left[T_k = t_{i,k}, T'_k = t'_{i,k} | \bar{t}_{i,k-1}, \overline{t'}_{i,k-1}, \bar{c}_{i,k}\right] \\ &= \prod_{k=0}^{n} \mathrm{P}\left[T'_k = t'_{i,k} | t_{i,k}, \bar{t}_{i,k-1}, \overline{t'}_{i,k-1}, \bar{c}_{i,k}\right] \cdot \mathrm{P}\left[T_k = t_{i,k} | \bar{t}_{i,k-1}, \overline{t'}_{i,k-1}, \bar{c}_{i,k}\right] \\ &= \prod_{k=0}^{K} \mathrm{P}\left[T'_k = t'_{i,k} | \bar{t}_{i,k}, \overline{t'}_{i,k-1}, \bar{c}_{i,k}\right] \cdot \mathrm{P}\left[T_k = t_{i,k} | \bar{t}_{i,k-1}, \overline{t'}_{i,k-1}, \bar{c}_{i,k}\right] \\ &= \left(w_{i,T} \cdot w^{\dagger}_{i,T'}\right)^{-1}. \end{aligned}$$

Where $w_{i,T}$ is the weight of the first treatment as defined originally and $w^{\dagger}_{i,T'}$ is the weight for the second treatment where we not only condition on the history of the confounder and the history of the second treatment, but also on the history of the first treatment.

In this manner, we can even add more time dependent variables and just multiply all the weights. In doing so the dependency structure between the weights does not matter, as long as no future treatment influence any past treatment and the outcome does not have a influence on any of the treatments.

A special application of these multiple treatments is so called censoring. In many survival studies the time of death is not known for all the patients because the study has to stop at one point and not all of the patients are dead yet. In this case we are actually interested in the survival given the intervention $do(t_0, \ldots, t_n)$ and given no censoring will occur. So we can define the time-dependent random variable $T'(t)$ to be the indicator variable for censoring at time t.

In the following we will refer to variables that fulfil the structure of T and T' in Graph 3.3 as generalized treatments. Generalized because variables like censoring are not treatments in the common sense.

Chapter 4

Survival Analysis

In this chapter we are briefly going to recall some theory about survival analysis with the Cox proportional hazards model and the extended Cox model with time-dependent covariates.

Finally we are going to see how to implement these survival models in R.

Note that in the appendix we will explain pooled logistic regression, which is an approximation of the cox proportional hazards model.

4.1 Introduction

When we are interested in the time until some failure occurs, we make a so called survival analysis. We might want to compare the expected survival for two or more groups being exposed to different circumstances.

One example for exposure is the treatment with some drug and we are interested in how much longer a treated patient is expected to live compared to a patient who was not treated. Other examples could be the life expectancy of a car if it is exposed to the elements rather than parked in a garage at night, the life expectancy of a smoker compared to a non-smoker or the time to the cure of some illness of two cohorts treated with different drugs.

Note that if the failure times of all the objects in the study were known, we could just fit a generalized linear model for the survival time. In most studies in real life however, the exact time of failure is not known for all objects. The reason for that is mainly that the study ends before the failure of all patients is detected. All the objects with no observed failure are said to be right censored, and it would bias the analysis if we would just exclude the objects with no observed failure in our analysis. Clearly the censored objects also give us some information about the survival times, particularly that they will survive longer than their observed follow-up time.

This is why we need a special approach to analyse censored data.

What we are basically going to do is to count the number of patients at risk at each time period and compare it to the number of patients failing at that time period. A patient is said to be at risk at time t if his or her total follow-up time is greater than or equal to t.

In a survival analysis there are mainly two functions of interest. Let T denote the random variable denoting the failing time, then we have:

- Survivor function: $S(t) = \mathrm{P}\left[T > t\right]$
- Hazard function: $h(t) = lim_{\Delta t \to 0} \frac{\mathrm{P}[t \leq T < t + \Delta t | T \geq t]}{\Delta t}$.

Note that the survivor function is always a decreasing function, while the hazard can be an arbitrary function in $\mathbb{R}^+$.

The hazard function represents the instantaneous potential of failing, i.e. given a person did not fail until time t, the hazard of t gives the potential of failing in the next time period. Note also that the hazard is not a probability, it can be greater than one.

It is possible to derive either of these two functions by knowing the other via:

- $S(t) = \exp\left(-\int_0^t h(u)du\right)$
- $h(t) = -\frac{S'(t)}{S(t)}$

In reality the time often is a discrete variable. Assume we have observations at times $t_1, \ldots t_n$, then we have the following useful decomposition of the survivor function:

$$\begin{aligned} S(t_i) &= \mathrm{P}\left[T > t_i\right] \\ &= \mathrm{P}\left[T > t_i | T > t_{i-1}\right] \cdot \mathrm{P}\left[T > t_{i-1}\right] \\ &= \ldots \\ &= \prod_{k=1}^{i} \mathrm{P}\left[T > t_k | T > t_{k-1}\right], (\ t_0 = 0) \\ &= \prod_{k=1}^{i} (1 - \mathrm{P}\left[t_k \leq T < t_{k+1} | T \geq t_k\right]) \\ &= \prod_{k=1}^{i} (1 - (t_{k+1} - t_k) \cdot h(t_k)). \end{aligned} \tag{4.1.0.1}$$

This follows directly from the definition of the conditional probability $\mathrm{P}\left[A|B\right] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]}$ with $A = \{T > t_i\}$ and $B = \{T > t_{i-1}\}$ so that $A \cap B = A$.

4.2 The Cox Proportional Hazards Model

Let us denote the exposure (i.e. treatment) with the random variable X_0.

Among others, a possible model for the hazard is the so called Cox proportional hazards model. It is preferential over many other models because it is only semi-parametric. We will only discuss the Cox proportional hazards model, for more information about the other models see for example Kleinbaum and Klein (2005), and for more information about survival analysis in general, see for example Therneau and Grambsch (2000).

Let us assume the survival is dependent on the set of random variables $(X_1, \ldots, X_n)$.

The Cox proportional hazards model assumes the following structure of the hazard function:

$$h(t) = h_0(t) \cdot \exp(\overline{\beta}^T \cdot \overline{X}), \tag{4.2.0.2}$$

where $h_0(t)$ is the so called baseline hazard, $\overline{X}$ is the vector of the covariates and $\overline{\beta}$ is a vector of parameters that have to be fit.

Note that the first term on the right hand side of (4.2.0.2) is only dependent on the time while the second term is not dependent on the time.

This implies that the hazard ratio

$$HR(\overline{X}_1, \overline{X}_2) := \frac{h_1(t)}{h_2(t)} = \exp(\overline{\beta}^T \cdot (\overline{X}_1 - \overline{X}_2)),$$

for two patients with covariate values $\overline{X}_1$ and $\overline{X}_2$ is not dependent on the time ($h_i(t)$ denotes the hazard for the patient with covariate values X_i). This is the so called proportional hazards assumption.

The measure of effect in a survival analysis is the hazard ratio of two different covariate values. When we refer to the hazard ratio for the covariate X_1, we mean that we take the hazard ratio of two patients that have the same covariate values for all the variables but X_1, i.e.:

$$HR(\overline{X}_1, \overline{X}_2) = \exp(\sum_{k=1}^{n} \beta_k \cdot (X_{k,1} - X_{k,2})) = e^{(X_{1,1} - X_{1,2}) \cdot \beta_1},$$

where $X_{1,1}$ and $X_{1,2}$ are the values of X_1 for patient one and two. If we want the hazard ratio for a dichotomous exposure variable ($X_1 = 1$ if exposed, $X_2 = 0$ if not exposed) the hazard ratio becomes e^{β_1}.

For a dichotomous exposure variable we can draw the following conclusions from the HR:

- If $HR = \alpha > 1$ (i.e. $\beta_1 > 0$) then the exposed group has a α times bigger hazard of dying.
- If $HR = 1$ (i.e. $\beta_1 = 0$) then the exposed has no effect on the survival.
- If $HR = \alpha < 1$ (i.e. $\beta_1 < 0$) then the unexposed group has a α^{-1} times bigger hazard of dying.

In summary this results in:

$$\beta_1 < 0 \Rightarrow \text{ exposure is successful.}$$

4.3 Extension of the Cox Model

So far we only had time independent covariates in our model, but clearly there could also be time dependent variables. One example of that is when the treatment is time dependent. In this case a patient is not either on treatment or not, but his treatment regime changes over time. Consider again Example 3.2.1 from chapter two. A patient enters the study at some date and the blood pressure is monitored at several follow-up times. Then the doctors decide when to start and stop treatment with a beta-blocker according to the blood pressure and possibly according to some other unmeasured factors

like personal preferences. Hence the exposure variable and the blood pressure are time dependent covariates, that possibly change at each follow-up.

For cases like this we are going to need an extension of the Cox proportional hazards model, called the extended Cox model. Let $\{X_1, \dots, X_{n_1}\}$ be the time independent covariates and $\{X_{n_1+1}, \dots, X_n\}$ the time dependent covariates. Then the the model of the hazard becomes:

$$h(t) = h_0(t) \cdot \exp\Big(\sum_{i=1}^{n_1} \beta_i \cdot X_i + \sum_{i=n_1+1}^{n} \beta_i \cdot X_i(t)\Big).$$

Clearly, for the hazard in this form we do not require the proportional hazards assumption. The assumption we make here is that the hazard at time t is only dependent on the covariate level at time t. In reality this assumption generally does not hold. An easy solution to this problem is to introduce some variables representing values of the covariate history that we consider important to adjust for. This could be the treatment status at time $t-1$, the total treatment dose a patient received so far or the like.

We are going to make an easy example where only the exposure variable is time-dependent and give a possible interpretation of the estimated parameters. But note that in presence of multiple time-dependent covariates it is more difficult to have a meaningful interpretation of the coefficients.

Example 4.3.1. Consider the setting from Example 3.2.1 and assume the baseline blood pressure is measured and the patient i starts taking the beta-blocker at some time t_i after the begin of the study. A patient is assumed to stay on treatment once he or she started. The random variable for the treatment is now a vector with entries zero for $X(t_1), \dots X(t_{i-1})$ and one afterwards. Table 4.1 shows hypothetical results of the survival analysis with the extended Cox model.

	coef	exp(coef)
treatment T	-0.7	0.5
baseline blood pressure B	1.1	3

Table 4.1: Hypothetical results for the extended Cox model of the heart attack example.

This implies that the hazard is:

$$h(t) = h_0(t) \cdot \exp\left(-0.7 \cdot T(t) + 1.1 \cdot B\right),$$

where $T(t)$ is the indicator variable for the treatment at time t and B the baseline blood pressure.
The hazard ratio at time t for two patients with the same baseline value is then:

$$HR(X_1(t), X_2(t)) = \exp\Big(-0.7 \cdot (T_1(t) - T_2(t))\Big).$$

Assume that patient one starts treatment at time t_1 and patient two at time $t_2 > t_1$, then the HR becomes:

$$HR(X_1(t), X_2(t)) = \begin{cases} 1 & \text{if t} \notin [t_1, t_2) \\ 0.5 & \text{if t} \in [t_1, t_2), \end{cases}$$

implying that only during the period $[t_1, t_2)$ the hazard of the untreated patient is twice as large as the hazard of the treated.

In this example we also see the impact of the assumption we made in the extended Cox model, namely that the hazard of the two patients is the same after the latter also started treatment. In reality it is not always realistic to assume that the patient that started later with the treatment should have the same hazard as the patient who started earlier, after both started treatment. One could for example assume that the treatment status one month before, or the total duration of treatment influences the hazard at present time.

Note that the fact that the hazard of the two patients is the same after t_2 does not imply that the survival curve is also the same. With (4.1.0.1) we can easily see that the survivor curve is influenced by all the subsequent hazards.

Another way to interpret the hazard ratio in this case is to treat $X_1(t)$ and $X_2(t)$ as the two possible future treatments for a patient. In this case the patient would have a twice as large hazard in the next time period if he remains untreated as if he were to start treatment in the mean time.

4.4 Implementation in R

In R, time dependent variables in a survival analysis are addressed via the counting process. For information about the programs Stata, SAS and SPSS see for example Kleinbaum and Klein (2005).

We first have to get the data in the right form to be able to execute the analysis. That is, for each time point where there is a new measurement of the time dependent variables we have to add a line to the data, including the start and end time of the time period, all the covariate values and an indicator for the failure. So there are going to be multiple rows per person.

As an example consider a patient with follow-ups at the fifth, sixth and ninth month after the first visit and blood pressure measurements of 120, 140, 130, 140 at each follow-up respectively. The patient started treatment after the second follow-up (month five) and his death was detected in the twelfth month after the first visit, so there was no more information about the patient after the 13th month. Then this patient contributes 4 rows to the data as shown in Figure 4.2. Note that the time intervals are closed on the left and open on the right, i.e. in the interval $[6, 9)$ the blood pressure is assumed to be 130 (over the whole interval), the patient was treated with beta blocker and not yet dead shortly before month 9.

patient	start	end	blood pressure	beta blocker	death
1	0	5	120	0	0
1	5	6	140	1	0
1	6	9	130	1	0
1	9	13	140	1	1
2	...				
⋮					

Table 4.2: Data layout for the analysis of time-dependent covariates in survival analysis with the Cox Model.

For this Cox model, we are going to use the commands of the `survival` package. The most important commands are:

```
formula <- Surv(start, end, death)
model <- coxph(formula ~ blood pressure + beta blocker)
```

For the Cox proportional hazards model (without time-dependent covariates) we only have to specify the first time argument which then stands for the total duration a patient contributed to the study. If `end` is also specified, then the `coxph` function automatically uses the counting process approach. Note that the same formula can also be used for a setting where multiple events are possible.

The summary of the model, among others, includes the coefficients of all the dependent variables and some test statistics. If we are interested in the survivor curve, we can use the following command:

```
survivor <- survfit(model, data=data[patient=i,], individual=TRUE)
s <- summary(survivor)
```

The output of the summary of the survfit function includes the survivor curve of the patient i. That is an array of all the hazard times and the corresponding values of the survivor function.

If we would omit the last argument of the survfit function, each line of data would be treated as a separate patient and the command would return a separate survival curve for each of them.

Note that in the output there are only values for the survivor at times where at least one failure is detected. With Formula (4.1.0.1), we see that we always multiply the former survivor value with the probability of surviving the next time period given survived the last time period. In case of no failure this probability is one and therefore the value of the survivor function is the same as one time point before. This is independent on the values of the covariates, meaning that if no one failed in some time period, then we simply assume that the probability of failing in this time period is zero, independent on the values of the covariates.

A possible way of receiving the survivor values for each time point of each patient is:

```
survivor.complete <- rep(NA,max(end))
survivor.complete[1] <- 1
survivor.complete[s$time] <- s$surv
survivor.complete <- na.locf(survivor.complete)
```

We use the command `na.locf` (last observation carried forward) which is included in the `zoo` package.

Chapter 5

Analysis of the Data

5.1 Description of the Data Set

The data is collected from 80488 HIV patients in Malawi and Zambia. For our analysis we only use the data of the 2658 patients that experienced immunologic failure and have no missing values in the baseline covariates. Starting time is the time of immunologic failure and the end time is either death, right censoring or lost to follow up. Every patient contributes multiple rows to the data. We use monthly data, but note that we could also use any other format for the time, like quarterly data for example. The range of the time period of a single patient is from one to 59 months with median 13 months and mean 16.5 months.

For each person the CD4 count was detected at several follow-up times. For months where there was no new information about the CD4 count, the last known value is assigned.

There is an indicator variable for death, and also an indicator variable for right censoring and lost to follow up respectively. These three variables are zero for all the rows except the last row of each patient. In the last row there is a one on exactly one of these three indicator variables. Note the difference between right censored and lost to follow-up is that a patient who is right censored is known to be alive at the end of the study, while a patient who is lost to follow-up did not come back to the next follow-up and we do not know if he or she is dead or alive. In our data there were 141 detected deaths, 2388 were right censored and 129 patients lost to follow-up.

Moreover we have a indicator variable for the switch to second line, we assume that a patient stays on second line treatment after starting. Hence this variable is one at the first month of the use of second line treatment and remain one until death, censoring or lost to follow-up occurs. There is a total of 376 patients being switched to second line treatment.

There are several baseline covariates, with a fixed value for each patient not changing over time. In particular we have the basic age group of the subject (< 30, 30-39, ≥ 40), the gender, an indicator if the patient was in advanced clinical stage at baseline and the basic CD4 count group ($<$49, 50-99, 100-199, ≥ 200).

All the patients with missing values in the baseline covariates are excluded and all the missing values of the CD4 count are assigned values as described above. Hence there are no missing values in the data set.

In Table 5.1 the distribution of all the patients into the groups is illustrated together with the ratios of death.

	total	**death**	**death rate**
Total	2658	141	5.30 %
Gender			
Female	1244	71	5.71 %
Male	1414	70	4.95 %
Age at Baseline			
<30	1751	92	5.25 %
30-39	747	41	5.49 %
≥ 40	16	8	50.00 %
CD4 at Baseline			
<50	361	40	11.08 %
50-99	635	50	7.87 %
100-199	844	37	4.38 %
≥ 200	818	14	1.71 %
Advanced stage			
No	715	28	3.92 %
Yes	1943	113	5.82 %
Second line			
No	2282	132	5.78 %
Yes	376	9	2.39 %

Table 5.1: Frequency of death in groups of the baseline variables.

In this overview of the data we can see that the death rate is smaller for the patients who are switched to second line treatment compared to those who stay on first line treatment after immunologic failure. But we have to be careful interpreting (causally) the ratios of this table, since they do not account for the fact that the treatment decision might be influenced by the CD4 count.

Notation: In this chapter we are going to use the following notations:

- $Y(t)$ = Indicator of death at time t (1=dead, 0=alive)
- $S = \min\{t; Y(t) = 1\}$ Time from start to death (∞ if no death detected)
- $T(t)$ = Indicator variable of second line treatment at time t (0=first line, 1= second line)
- $SL = \min\{t; T(t) = 1\}$ Time from start to start of second line treatment (∞ if never switched)
- $T'(t)$ = Indicator variable of censoring at time t (0=uncensored, 1=censored)
- $Cens = \min\{t; T'(t) = 1\}$ Time from start to censoring (∞ if not censored)
- $T''(t)$ = Indicator variable of loss to follow-up at time t (0=not lost, 1=lost)
- $Loss = \min\{t; T''(t) = 1\}$ Time from start to loss to follow-up (∞ if not lost)
- $C(t)$ = CD4 count measurement at time t

- $\overline{B}$ = Vector of baseline covariates.
- We call $[j-1, j)$ the j-th month and the start 0 is the date of immunologic failure.
- The over-bar of a time dependent variable denotes the (discrete) history of it, for example $\overline{C(t)} = \{C(0), C(1), \ldots, C(t)\}$.

Data layout: Each person contributes multiple rows to our data, each representing a month of his or her total follow-up time. As a (fictitious) example consider a person who is switched to second line treatment in the third month after immunological failure and at the fifth month after immunological failure the study ends and we know the patient is still alive. This would result in the data given in Table 5.2.

Patient	Month	$C(t)$	$T(t)$	Y	$T'(t)$	$T''(t)$	$\overline{B}$
1	1	*	0	0	0	0	*
1	2	*	0	0	0	0	*
1	3	*	1	0	0	0	*
1	4	*	1	0	0	0	*
1	5	*	1	0	1	0	*
2	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 5.2: Data of a fictitious HIV patient.

5.2 Analysis of the Data

The ideal randomized trial for this set-up would be to assign second line treatment randomly to 50% of the patients and make them switch immediately after immunologic failure. Obviously it is not feasible to do so because of ethical reasons and therefore it is important that we can also draw conclusions from the observational study.

We want to know the effect of the treatment regime on the survival. I.e we want to know the hazard ratio of patients who are switched and patients who are not. It is likely that the decision to start second line treatment is dependent on the CD4 count. But a low CD4 count also implies a more severe state of the disease and therefore a shorter life expectancy. Hence we have a time dependent confounder which is also a intermediate variable for treatment and the outcome.

As mentioned in previous chapters the standard approach of dealing with survival times will be biased in this data set due to this confounding and we will have to apply MSM to receive unbiased estimates.

In Appendix B the outputs of the R computation are given for the most important models.

Notation: In the following we are going to write

$$T \sim X_1(t) + X_2(t) + \ldots + X_n(t)$$

for the survival model:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\mathrm{P}\left[t \leq T < t + \Delta t | T \geq t, \overline{X}(t)\right]}{\Delta t} \\ &= h_0(t) \cdot \exp(\alpha_1 \cdot X_1(t) + \ldots + \alpha_n \cdot X_n(t)) \end{aligned}$$

5.2.1 Crude Analysis

We first want to do a crude analysis of the data to see the impact of the circumstance that the CD4 count is a confounder as well as an intermediate variable. Clearly we have to include all the baseline covariates for the survivor function. The only question is if we want to include the CD4 count or not.

There is no 'right' answer to this question, since we are aware of the fact that the CD4 count is a confounder and an intermediate at the same time. Therefore we fit the two models:

$$S \sim T(t) + \overline{B}$$
$$S \sim C(t) + T(t) + \overline{B}.$$

It turns out that the coefficient for the second line treatment do not differ much in these two models. The coefficient is -0.922 for the first models and -0.944 for the second model, this results in a hazard ratio of 0.398 (95% CI (0.20-0.79)) and 0.389 (95% CI (0.20-0.77)). The interpretation of the coefficient is the following: Comparing two patients with the same baseline values and the same CD4 history, where one patient is on second line treatment and one not (in this specific time point), then the treated has about 0.4 times the hazard of the untreated.

Note that the small difference of the coefficients of the two models does not necessary indicate a small influence of CD4 on second line treatment. We can see in Figure 5.1 that the CD4 count indeed influences the decision to switch treatment. In the left panel of Figure 5.1 we included all the observations up to the first initiation of second line treatment. In the right panel we included only the patients who switch to second line treatment, again only up to first observation of initiation of second line treatment. The dotted line is the estimated probability of treatment from the simple logistic model.

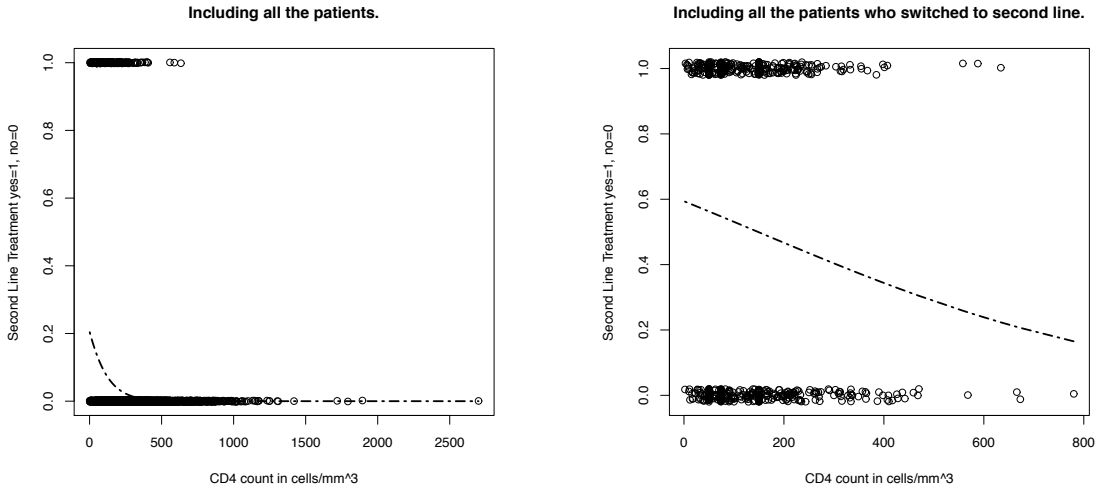

Figure 5.1: Plot of the start of second line treatment on the CD4 count. In the left panel all the patients are included, in the right panel only those who had a switch.

More precisely, fitting the model $SL \sim C(t) + \overline{B}$ results in a coefficient of -0.0085 for the CD4 count. If we assume an increase of the CD4 count of about 100 cells per mm^3 then

the hazard ratio is 0.427, meaning that a patient with 100 cells more than another patient has about 0.4 times the hazard of receiving treatment.

5.2.2 IPTW Weighting with a Simple Model

We want to weight the whole population in order to receive a new data set in which the switching of the second line is no longer dependent on the CD4 count. Note that since we treat each person month as an observation we also have to weight each person month separately. We have three kind of generalized treatments: the second line treatment T, the censoring T' and the lost to follow-up T''.

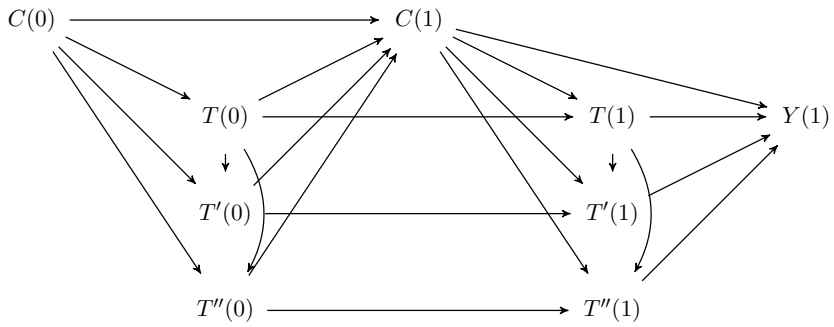

Figure 5.2: Assumed graph for our analysis with two time points. T denotes the second line treatment, T' the censoring and T'' the loss to follow-up.

The assumed graph is shown in Figure 5.2. We assume that the treatment can influence the censoring and the loss to follow-up, but censoring and follow-up do not have an influence on the treatment. Note that similar to the figure in Chapter 3 (Figure 3.3) we omit some arrows for the sake of simplicity. Whenever there is a directed path between two knots, then this also stands for a direct arrow between these knots.

Remember from Section 3.4 that the weight for patient i at month j in this case is:

$$
\begin{aligned}
w_i(j)^{-1} = &\prod_{k=1}^{j} \mathrm{P}\left[T(k) = t_i(k) | \overline{(t_i(k-1))}, \overline{(t_i'(k-1))}, \overline{(t_i''(k-1))}, \overline{c(k)}, \bar{b}\right] \\
&\cdot \prod_{k=1}^{j} \mathrm{P}\left[T'(k) = t_i'(k) | \overline{(t_i(k))}, \overline{(t_i'(k-1))}, \overline{(t_i''(k-1))}, \overline{c(k)}, \bar{b}\right] \\
&\cdot \prod_{k=1}^{j} \mathrm{P}\left[T''(k) = t_i'(k) | \overline{(t_i(k))}, \overline{(t_i'(k))}, \overline{(t_i''(k-1))}, \overline{c(k)}, \bar{b}\right] \\
= &\prod_{k=1}^{j} \mathrm{P}\left[T(k) = t_i(k) | \overline{(t_i(k-1))}, \overline{c(k)}, \bar{b}\right] \\
&\cdot \prod_{k=1}^{j} \mathrm{P}\left[T'(k) = t_i'(k) | \overline{(t_i(k))}, \overline{(t_i'(k-1))}, \overline{c(k)}, \bar{b}\right] \\
&\cdot \prod_{k=1}^{j} \mathrm{P}\left[T''(k) = t_i'(k) | \overline{(t_i(k))}, \overline{(t_i''(k-1))}, \overline{c(k)}, \bar{b}\right] \\
:= &w_{i,T}(j)^{-1} \cdot w_{i,T'}^{\dagger}(j)^{-1} \cdot w_{i,T''}^{\dagger}(j)^{-1},
\end{aligned}
$$

where $w_{i,T}$ is the weight for the second line treatment as originally defined and $w^{\dagger}_{i,T'}$ is the weight for censoring where we not only condition on the confounder history and the history of T' but also on the history of T.

The second equation follows from the fact that $T(k)$ and $\overline{T'(k-1)}$ are conditionally independent given $\overline{C(k)}$. The same argument applies also on $T(k)$ and $\overline{T''(k-1)}$, $T'(k)$ and $\overline{T''(k-1)}$, $T''(k)$ and $\overline{T'(k)}$.

Throughout the whole analysis we use the stabilized weights but for simplicity denote them with 'w' rather than 'sw'. The equation for the weights above applies to the stabilized weights in the exact same way.

We are first going to explain the method with a basic model for S, SL, $Cens$ and $Loss$. Later we are going to try different model in order to determine the best suitable model, and the sensitivity of the results to model choice.

As mentioned in the previous chapter, the assumption that the hazard of treatment at a given time-point is only dependent on the CD4 count at that same time-point is likely to be violated in our set-up. Therefore we include a so called lag variable of the CD4 count in the analysis, which is defined as $C.lag1(j) = C(j-1)$.

Weights for Second Line Treatment

We first want to derive $w_{i,T}(j)$ for all the patient i and months j. Note that in our study a patient stays on second line treatment forever when once started, hence

$$\mathrm{P}\left[T(j) = 1 | T(j-1) = 1\right] = 1,$$

independent on the value of $\overline{T(j-2)}$ and $\overline{B}$. So we only need to fit a model for the part of the data up to the first initiation of second line treatment. One can easily see that under this assumption our task is actually to fit a survival model where the start of second line treatment is the failure. So we get the following formulas for the weights:

- For j so that $t_i(j) = 0$ the weights simplify to:

$$w_{i,T}(j) = \frac{\mathrm{P}\left[SL > j | SL > j-1, \overline{B} = \overline{b}_i)\right]}{\mathrm{P}\left[SL > j | SL > j-1, \overline{B} = \overline{b}_i, \overline{C(j)} = \overline{c_i(j)}\right]},$$

  with Equation (4.1.0.1).

- For the weights $w_{i,T}(j)$ with $t_i(j) = 1$, let $k = \min\{j | t_i(j) = 1\}$ be the first month

of second line treatment. Then we can decompose the product:

$$\begin{aligned}
&\prod_{l=1}^{j} \mathrm{P}\left[T(l) = t(l)|\overline{T(l-1)} = \overline{t_i(l-1)}\right] \\
&= \Big(\prod_{l=1}^{k-1} \mathrm{P}\left[T(l) = 0|T(l-1) = 0\right]\Big) \cdot \Big(\mathrm{P}\left[T(k) = 1|T(k-1) = 0\right]\Big) \\
&\quad \cdot \Big(\prod_{l=k+1}^{j} \underbrace{\mathrm{P}\left[T(l) = 1|T(l-1) = 1\right]}_{1}\Big) \\
&= \mathrm{P}\left[SL > k-1\right] \cdot (1 - \mathrm{P}\left[T(k) = 0|T(k-1) = 0\right]) \\
&= \mathrm{P}\left[SL > k-1\right] \cdot (1 - \mathrm{P}\left[SL > k|SL > k-1\right]) \\
&= \mathrm{P}\left[SL > k-1\right] - \mathrm{P}\left[SL > k\right].
\end{aligned}$$

Hence the weights $w_{i,T}(k), w_{i,T}(k+1), \ldots, w_{i,T}(t.end_i)$ are all equal and calculated as:

$$w_i(j) = \frac{\mathrm{P}\left[SL > k-1|\overline{B} = \bar{b}_i\right] - \mathrm{P}\left[SL > k|\overline{B} = \bar{b}_i\right]}{\mathrm{P}\left[SL > k-1|\overline{B} = \bar{b}_i, \overline{c_i(k-1)}\right] - \mathrm{P}\left[SL > k|\overline{B} = \bar{b}_i, \overline{c_i(k)}\right]},$$

We can clearly not include the whole CD4 count history $\overline{C(j)}$ into the model, in our analysis we include the two variables $C(j)$ and $C.lag1(j)$.

Even though we do not believe that this restriction captures the circumstance completely adequate, we at least believe that the omitted variables, compared to the two we include, do not have a large impact on the treatment decision.

So the two models would be:

$$\begin{aligned}
SL &\sim \overline{B} \\
SL &\sim \overline{B} + C(t) + C.lag1(t),
\end{aligned}$$

for the numerator and denominator of the stabilized weights respectively.

Weights for the Censoring and Lost to follow-up

Censoring and Lost to follow-up are both typical cases for survival analysis. They both happen once and there is no more information after the first failure. So the weights are calculated the same way as the weights for the treatment, with the only difference that there is at most one weight to be calculated with the second equation from above.
For the model specification we use the same models except that we include the second line treatment discussed earlier in this chapter.

The two models for censoring are then:

$$\begin{aligned}
Cens &\sim T(t) + \overline{B} \\
Cens &\sim T(t) + \overline{B} + C(t) + C.lag1(t),
\end{aligned}$$

and the models for loss to follow-up:

$$\begin{aligned}
Loss &\sim T(t) + \overline{B} \\
Loss &\sim T(t) + \overline{B} + C(t) + C.lag1(t).
\end{aligned}$$

Analysis with the Weights

By weighting the population with the product of the three weights we assure that the treatment is no longer confounded by the CD4 count. Hence we can apply an extended Cox model to the weighted data without having the problem described in the crude analysis. The model we fit in the pseudo-population is then:

$$S \sim T(t) + \overline{B}$$

The resulting coefficient is -1.137, which results in a hazard ratio of 0.321(95% CI (0.15-0.68). Compared to the ratio of the crude estimation (0.398), we can see that the crude analysis indeed underestimates the effect, but only slightly.

5.2.3 Interpretation

So far we only interpreted the causal coefficient of the treatment, but clearly there are more coefficients to interpret, namely the coefficients for the baseline covariates. In the following we use the model excluding $C(t)$ and $C.lag1(t)$ as the crude model. The coefficients for the other model only differ slightly.
The baseline covariates are:

- Gender: 1 for male and 0 for female
- Age: Age group (<30, 30-39, $\geq$ 40) at the beginning of the study
- CD4: CD4 group (<50, 50-99, 100-199, $\geq$ 200) at the date of immunological failure
- Stage: Indicator if disease is in an advanced stage

The output of the weighted Cox regression together with the p-value from the likelihood ratio test is shown in Table 5.3.

	Crude Analysis			**Weighted Analysis**		
	Coefficient	Ratio	p-value	Coefficient	Ratio	p-value
Second Line			0.008			0.003
No	0 (reference)	1		0 (reference)	1	
Yes	-0.922	0.398		-1.137	0.321	
Gender			0.529		0.478	
Female	0 (reference)	1		0 (reference)	1	
Male	-0.108	0.898		-0.121	0.886	
Age at Baseline			0.962			0.933
<30	-0.023	0.977		-0.052	0.950	
30-39	0 (reference)	1		0 (reference)	1	
$\geq$ 40	-0.106	0.899		-0.125	0.883	
CD4 at Baseline			1.17e-11			1.50e-12
<50	1.963	7.120		2.014	7.493	
50-99	1.482	4.401		1.494	4.454	
100-199	0.937	2.553		0.935	2.547	
$\geq$ 200	0 (reference)	1		0 (reference)	1	
Advanced stage			0.109			0.090
No	-0.330	0.719		-0.347	0.707	
Yes	0 (reference)	1		0 (reference)	1	

Table 5.3: Coefficients of the fitted Cox model.

For completeness we give interpretations of all the point estimates, but one should keep in mind that only the coefficients of the baseline CD4 count is significantly different than one.

- The hazard ratio of a male and a female with the same treatment, age group, CD4 baseline group and Stage, is $e^{-0.108} = 0.898$, i.e. the survival is slightly better for men.
- The two negative coefficients for the age groups imply that the patients younger than 30 and older than 40 both have a better survival than the ones aged 30 to 39, i.e. they have hazard ratios of $e^{-0.023} = 0.977$ and $e^{-0.106} = 0.899$ compared to the group of age between 30 and 39. Hence the patients aged between 30 and 39 have the worst survival, and the ones above 40 have the best survival expectancy.
- As for the CD4 count the coefficients imply that the higher the value the better the survival. The three groups have a ratio of $e^{0.937} = 2.553$, $e^{1.482} = 4.401$ and $e^{1.963} = 7.120$ compared to the group with a CD4 count above 200.
- The coefficient for the stage implies that a patient who is not in advanced stage at the beginning of the study has a hazard ratio of $e^{-0.330} = 0.719$ compared with a patient in advanced stage when all other variables are the same. Hence the expected survival is better for patients not in advanced stage.

Note that the effect of CD4 and Stage on the survival is comprehensible while the effect of the age group and gender is not explicable with logical consideration, but since the effect is comparable small, we believe that this output is actually reasonable.

5.2.4 Effect of the Switch to Second Line for the Different Groups

So far we only looked at the overall effect of the switch of treatment. This means we assumed that the effect of $T(t)$ is the same in all groups in the baseline covariates. In Table 5.4 the effect of second line treatment is given for each group.

Together with the coefficient and the hazard ratio there is the 95% confidence interval and the p-value from the likelihood ratio test when we compare the basic model with the model including the interaction term with $T(t)$.

Note that these interaction terms are all non-significant, so we should interpret these ratios restrainedly. Moreover the confidence intervals of the overall effect of second line treatment indicate that the difference of the effect in the crude analysis and the weighted is not that big; they both have almost the same confidence intervals.

The interpretation of the coefficients for gender is that the switch of treatment has a more beneficial effect on women than on men. This results should not be mixed up with the results of Table 5.3, where we concluded that men have a better survival than women in general. Moreover we can see that patient in the baseline age group 30-39 experience the best advancement of the survival when switched to second line, while patients above age 40 experience almost no advancement. The hazard ratios for the different baseline CD4 groups display the circumstance that there is no patient that switched to second line treatment in the group '≥ 200' and died. Therefore the coefficient of this group cannot be fitted (i.e. the coefficient should be minus infinity). Because of this the two models

	Crude Analysis			Weighted Analysis		
	Coefficient	Ratio (0.95 CI)	p-value	Coefficient	Ratio (0.95 CI)	p-value
Overall effect	-0.92	0.40 (0.20-0.79)	0.008	-1.14	0.32 (0.15-0.68)	0.003
Gender			0.39			0.57
Female	-1.22	0.30 (0.11-0.82)		-1.34	0.26 (0.09-0.77)	
Male	-0.62	0.54 (0.22-1.35)		-0.91	0.40 (0.14-1.13)	
Age at Baseline			0.28			0.33
<30	-0.69	0.50 (0.23-1.09)		-0.91	0.40 (0.17-0.97)	
30-39	-2.03	0.13 (0.02-0.96)		-2.12	0.12 (0.02-0.84)	
≥ 40	-0.10	0.91 (0.11-7.38)		-0.16	0.85 (0.10-7.12)	
CD4 at Baseline			0.75			0.66
<50	-0.81	0.45 (0.17-1.14)		-0.94	0.39 (0.14-1.06)	
50-99	-1.32	0.27 (0.07-1.11)		-1.75	0.17 (0.03-0.98)	
100-199	-0.48	0.62 (0.15-2.58)		-0.66	0.52 (0.11-2.36)	
≥ 200	-14.23	0.00 (0-inf)		-15.18	0.00 (0-inf)	
Advanced stage			0.09			0.14
No	-0.05	0.95 (0.33-2.77)		-0.32	0.73 (0.23-2.32)	
Yes	-1.30	0.27 (0.11-0.67)		-1.50	0.22 (0.08-0.60)	

Table 5.4: Coefficients for the effect of second line in the different groups.

cannot be compared and the p-value is not really meaningful. Finally the coefficient of the advanced stage tell us something important. We can see that there is actually a huge difference of the effect of switching in these two groups. For patients in advanced stage it is extremely beneficial to switch to second line treatment, while for patients not in advanced clinical stage the change of treatment has almost no effect on the survival.

5.2.5 Other Measures of Effect of the Second Line Treatment

So far we only examined the effect of the switch to second line treatment assuming that the hazard ratio is constant over time. What we might also want to know is if the time a patient waited until he started treatment has an influence on the hazard, or if the duration of second line treatment has an influence on the hazard. We want to have a quick look at those two questions in this section.

First we need to implement two more variables. The first represents the time until a patient is switched, we call it $Wait$. This variable increases by one in each time interval until the patient starts second line treatment and then stays constant, for example a patient who switched treatment in the interval $[4, 5)$ (in the 5th month) has $Wait = (0, 1, 2, 3, 4, 4, 4, \ldots)$, i.e. he stayed on first line treatment for four more months after immunological failure before he was switched to second line treatment. The other variable we are going to need is the total duration on second line (call it Tot) which is defined to be the the cumulative sum of the vector SL, i.e. with the example above, this patient has $Tot = (0, 0, 0, 0, 1, 2, 3, \ldots)$.

Let us first have a look at the time a patient waits until he starts treatment. For that purpose we fit the survival model

$$S \sim T(t) + Wait(t) + \overline{B},$$

in the original data and the data of the pseudo-population and get the output of Table 5.5

Note that the large p-value and the coefficient being close to zero implies that the effect of the duration of waiting until the switch is negligible compared to the effect of switching in general. Nevertheless we are briefly going to explain how the coefficients can be interpreted.

	Crude Analysis		Weighted Analysis	
	coeff	p-value	coeff	p-value
SL	-0.903	0.010	-1.095	0.004
Wait	0.017	0.483	0.021	0.405

Table 5.5: Output of the crude and weighted analysis with the additional variable $Wait$.

The interpretation is a little bit more complicated than before, and we basically have to distinguish between two cases:

- The hazard ratio of two patients, both on second line and patient one switched n months earlier than patient two, is:
$$\exp(\beta_1(T_1(t) - T_2(t)) + \beta_2(Wait_1(t) - Wait_2(t))) = e^{-n\cdot\beta_2} = 0.983^n \text{ (crude)},$$
$$= 0.979^n \text{ (weighted)}$$
- The hazard ratio of two patients, the first started second line treatment n months ago, the second patient is not on second line yet, is:
$$\exp(\beta_1 \cdot (T_1(t) - T_2(t)) + \beta_2(Wait_1(t) - Wait_2(t))) = e^{\beta_1 - n\cdot\beta_2} = 0.405 \cdot 0.983^n \text{ (crude)},$$
$$= 0.335 \cdot 0.979^n \text{ (weighted)}$$

This implies that the switch is quite beneficial and it is slightly more beneficial to switch as soon after immunologic failure as possible.

Second, we have a look at the total time the patient stays on second line treatment. We fit the model

$$S \sim T(t) + Tot(t) + \overline{B},$$

for original data and data of pseudo-population and get the output given in Table 5.6.

	Crude Analysis		Weighted Analysis	
	coeff	p-value	coeff	p-value
SL	-1.193	0.033	-1.344	0.028
Tot	0.030	0.531	0.023	0.647

Table 5.6: Output of the crude and weighted analysis with the additional variable Tot.

Again, to interpret this we have to consider two different cases, but here again, keep in mind that the effect is non-significant.

- The hazard ratio of two patients, both on second line and patient one was on second line n months longer than patient two, is:
$$\exp(\beta_1(T_1(t) - T_2(t)) + \beta_2(Tot_1(t) - Tot_2(t))) = e^{n\cdot\beta_2} = 1.030^n \text{ (crude)},$$
$$= 1.024^n \text{ (weighted)}$$
- The hazard ratio of two patients, the first is on second line treatment for n months, the second patient is not on second line yet, is:
$$\exp(\beta_1 \cdot (T_1(t) - T_2(t)) + \beta_2(Tot_1(t) - Tot_2(t))) = e^{\beta_1 + n\cdot\beta_2} = 0.303 \cdot 1.030^n \text{ (crude)},$$
$$= 0.261 \cdot 1.024^n \text{ (weighted)}$$

This implies that the switch is quite beneficial but the beneficial effect decreases slightly, the longer the patient is on second line treatment.

5.3 Normalized Stabilized Weights

Xiao et al. (2010) proposed another definition for the weights to construct the data of a pseudo-population, the so called normalized stabilized weights:

$$nsw_i(t) = \frac{sw_i(t) \cdot N(t)}{\sum_{i \in R(t)} sw_i(t)},$$

where $R(t)$ is the risk set at t, i.e. all the data rows with $month = t$, and $N(t)$ is the number of observations in the corresponding risk set, i.e. the number of elements in $R(t)$. Note that in that same way we can also define the unstabilized normalized weights by replacing the stabilized weights $sw_i(t)$ with the unstabilized weights $w_i(T)$ in the definition above.

Xiao et al. (2010) stated in their paper that the unstabilized weights tend to increase with increasing time, and the stabilized weights tend to decrease with increasing time. The normalization defined above assures that the average of the weights within each risk set is one, i.e. the average of the weights are about constant over time. In Figure 5.3 the

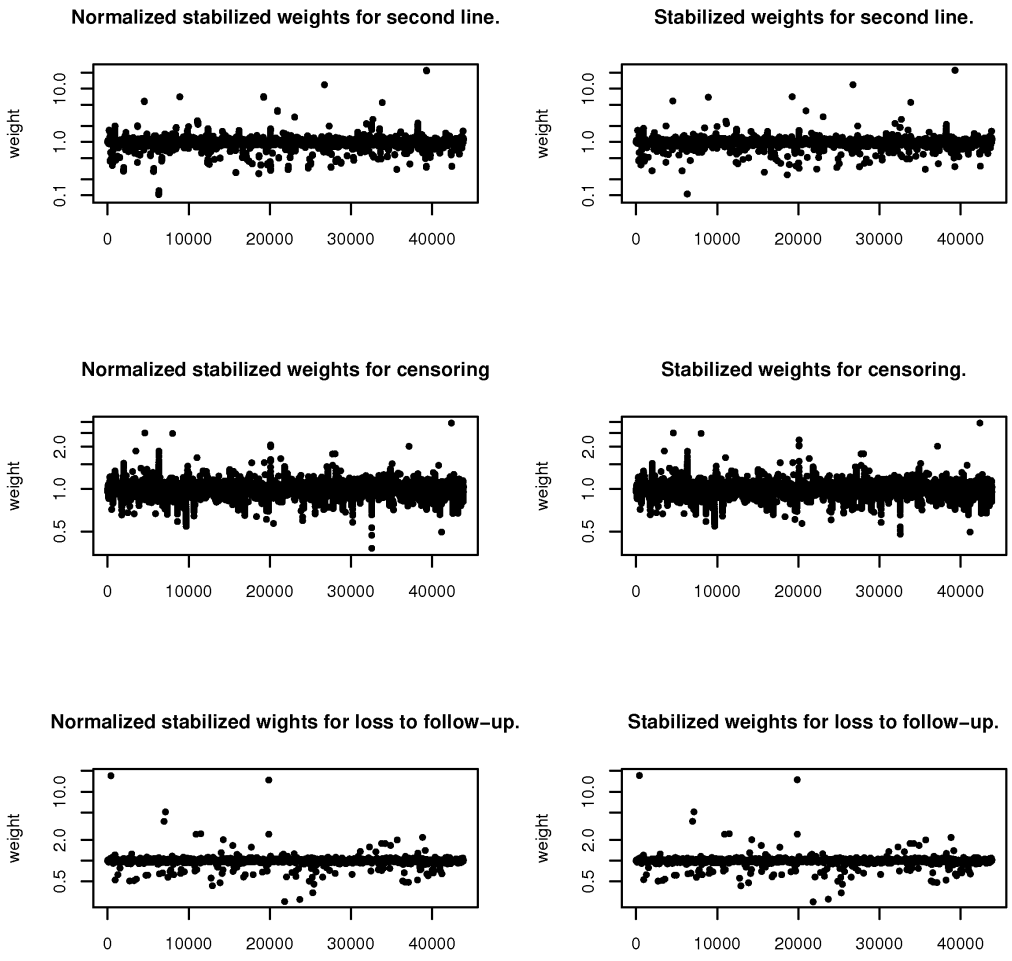

Figure 5.3: Normalized stabilized weights (left panel) and stabilized weights (right panel).

stabilized and the normalized stabilized weights are drawn for the second line treatment,

the censoring and the loss to follow-up. We can see that the weights are almost identical.

The result of the analysis with the normalized weights is that the second line treatment has a coefficient of -1.141 and a hazard ratio of 0.319, which is only slightly different from the analysis with the stabilized weights. The confidence interval and the p-value only differ by a thousandth.

So with our data, the normalization of the weights does not improve the results, but nevertheless it verifies the results we derived so far.

5.4 Sensitivity to Model Selection

In general we need to fit four models for the analysis of the data set. A model for the switch to second line treatment, a model for censoring, a model for loss to follow-up and a model for the survival.

The model from Chapter 5.2 is only the most simplest one, but other terms could be important to include in the analysis too, in particular interaction terms and covariates describing the course of the two time-dependent variables $T(t)$ and $C(t)$.

We already included the $C.lag1(t)$. There are a lot more possible functions of the CD4 count one could include, but we restrict ourselves to the CD4 count at current time and one month before. Moreover we include the total duration of second line treatment $Tot(t)$ defined in Chapter 5.2.5.

Notation: In the following we are going to illustrate the models with matrices as in Table 5.7.

	1	X_1	X_2
1		×	×
X_1	×		×
X_2	×	×	

Table 5.7: Example of a Model.

The first row and column represent the main effects. In this example we have that both variables X_1 and X_2 are in the model. The top left square stands for the intercept. When we use the extended Cox model, no intercept is needed, because is is incorporated in the baseline hazard. All the other squares stand for the interaction terms. If the example above were the model for Y, then together with the notation for the survival we established earlier, the Table 5.7 reads as:

$$Y \sim X_1 + X_2 + X_1 \cdot X_2$$

5.4.1 Models for the weights

In the following we are going to select three models for the treatment, censoring and lost to follow-up. The first is chosen with the AIC step backward criterion, the second model with the AIC step forward criterion and the third is the basic model from Chapter 5.2. We then do the IPTW analysis with all possible combinations of these models.

For these three cases we are going to define a maximal model including all the reasonable covariates and interaction terms, and a minimal model including the covariates we believe have to be in the model even if the criteria are not fulfilled. The maximal models are given in Table 5.8.

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×	×	×	×	×		
CD4	×			×	×		×		
CD4 lag1	×			×	×		×		
Gender	×	×	×		×	×	×		
Age	×	×	×	×		×	×		
Base CD4	×			×	×		×		
Stage	×	×	×	×	×	×			
SL									
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×	×	×	×	×	×	×
CD4	×			×	×		×	×	×
CD4 lag1	×			×	×		×	×	×
Gender	×	×	×		×	×	×	×	×
Age	×	×	×	×		×	×	×	×
Base CD4	×			×	×		×	×	×
Stage	×	×	×	×	×	×		×	×
SL	×	×	×	×	×	×	×		
Total SL	×	×	×	×	×	×	×		

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×	×	×	×	×	×	×
CD4	×			×	×		×	×	×
CD4 lag1	×			×	×		×	×	×
Gender	×	×	×			×	×	×	×
Age	×	×	×				×		
Base CD4	×			×			×		
Stage	×	×	×	×	×	×		×	×
SL	×	×	×	×			×		
Total SL	×	×	×	×			×		

Table 5.8: Maximal model for second line treatment, the censoring and the lost to follow-up.

The reason for not including the interaction terms of CD4 count, CD4 lag1 and baseline CD4 is that they are difficult to interpret. In the maximal model for the loss to follow-up there are six interaction terms less than in the maximal model for censoring. The reason is that in these six groups there are no observations, hence for these groups we cannot estimate a coefficient.

Note that in all the three cases we need a model for denominator and numerator simultaneously, that is, for the numerator we are just going to use the same model but reduced, i.e. excluding $C(t)$ and $C.lag1(t)$.

It turns out that the baseline CD4 count is highly significant in all the three reduced models, but not at all in the models for the denominator. Therefore the baseline CD4 count is one covariate we decide to include in the minimal model. Moreover the CD4 count at current time is the essential covariate that we want to control for, hence it is also included in the minimal model. Thus the minimal model is the same in the three cases and includes the baseline CD4 count and the CD4 count at current time.

Models for the Second Line Treatment

Table 5.9 gives the three models. Note that the CD4 count and the CD4 count at the previous month both have a negative coefficient. This means that the higher the CD4 count the smaller the odds of switching as we expected it.

Notes on unmeasured variables with influence: We do not really know on what the decision to switch to second line treatment really depends. According to the WHO criteria on the CD4 count all the patient should have been switched at the beginning of the study, but since this is not so,the switching seems somewhat arbitrary. An obvious confounder is the general health condition of a patient. Moreover, probable unmeasured variables of influence could be personal bias of the doctors, site of the medical center, availability of the second line treatment, wealth of the patient and so on.

Models for the Censoring

Table 5.10 gives the three models.

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×	×	×	×	×		
CD4	×			×	×				
CD4 lag1	×				×				
Gender	×	×							
Age	×	×	×						
Base CD4	×								
Stage	×								
SL									
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×			×	×		
CD4	×								
CD4 lag1	×								
Gender									
Age									
Base CD4	×								
Stage	×								
SL									
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×	×	×	×	×		
CD4	×								
CD4 lag1	×								
Gender	×								
Age	×								
Base CD4	×								
Stage	×								
SL									
Total SL									

Table 5.9: Best Models for second line treatment with AIC backward selection (left table),AIC forward selection (middle table) and the basic model (right table).

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×	×	×	×	×	×	×
CD4	×							×	×
CD4 lag1	×								
Gender	×					×		×	×
Age	×						×		
Base CD4	×			×				×	
Stage	×				×				
SL	×	×		×		×			
Total SL	×	×		×					

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×	×	×	×			×
CD4	×								×
CD4 lag1	×								
Gender	×					×			
Age	×								×
Base CD4	×			×					×
Stage									
SL									
Total SL	×	×			×	×			

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×	×	×	×	×	×	
CD4	×								
CD4 lag1	×								
Gender	×								
Age	×								
Base CD4	×								
Stage	×								
SL	×								
Total SL									

Table 5.10: Best Models for the censoring with AIC backward selection(left table), AIC forward selection (middle table) and the basic model (right table).

Notes on unmeasured variables with influence: Censoring basically correlates with better life expectancy, hence we could expect that only personal features of the patient influence the censoring. The most important covariates are already included in the baseline variables, but one more important variable could be the location of the patient.

Models for the Loss to Follow-up

Table 5.11 gives the three models. Note that the coefficient for the CD4 count is negative. This implies that the lower the CD4 count the more likely someone is lost to follow-up.

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×	×	×	×	×	×	
CD4	×			×			×		
CD4 lag1	×				×				
Gender	×	×					×	×	
Age	×		×						
Base CD4	×								
Stage	×	×		×					
SL	×			×					
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×			×	×		×	
CD4	×				×				
CD4 lag1									
Gender									
Age	×	×							
Base CD4	×								
Stage									
SL	×								
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1		×	×	×	×	×	×	×	
CD4	×								
CD4 lag1	×								
Gender	×								
Age	×								
Base CD4	×								
Stage	×								
SL	×								
Total SL									

Table 5.11: Best Models for the lost to follow-up with AIC backward selection (left table), AIC forward selection (middle table) and the basic model (right table).

Notes on unmeasured variables with influence: For the loss to follow-up we believe that there is a lot of influence coming from unmeasured variables. First there are features

of the patient itself influencing the lost to follow-up. Examples are: wealth, general health, location, relocation, confidence in the medic in charge, and so on. Second, the medical center can account for some more unmeasured confounder, features like preciseness with the collection of the data and how hard they try to contact a patient not coming back will clearly influence the amount of patients lost to follow-up.

Finally death itself is a main confounder. The fact that a low CD4 count increases the probability of being lost to follow-up is an indicator that sicker people are more likely to be lost to follow-up, and sicker people are also more likely to be dead.

5.4.2 Models for the Survival

In epidemiologic studies, the gender and age is always an important covariate, so it has to be in the minimal model. Also the switch to second line treatment has to be in the model because it is the goal variable. The maximal model is shown in Table 5.12. Clearly the time dependent CD4 count variables are not allowed to be in the model. Moreover we do not want to include the total duration of second line and any interaction terms with second line, because it makes it hard to interpret the effect of the switch.

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1				×	×			×	
CD4									
CD4 lag1									
Gender	×								
Age	×								
Base CD4									
Stage									
SL	×								
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1				×	×	×	×	×	
CD4									
CD4 lag1									
Gender	×				×	×	×		
Age	×			×		×	×		
Base CD4	×			×	×		×		
Stage	×			×	×	×			
SL	×								
Total SL									

Table 5.12: Minimal (left table) and maximal (right table) model of the weighted analysis for the survival.

The AIC forward and backward selection both result in the basic model. So in order to have three different models, we are going to use the second best models from the AIC backward and forward selections. The last variable that is dropped in the backwards selection is the interaction term of advanced stage and gender, so our fist model includes this interaction term together with all the main effects of the basic model.

The last variable that is added in the forward selection is the advanced stage, so the second model includes all the main effects of the basic models except the advanced stage. Note that the second model is the same model that we would obtain by the BIC stepwise selection. The three models for survival are given in Table 5.13.

5.4.3 Results

Impact of the Treatment Models

We first want to see the impact of the model choice for the weights on the output of the weighted survival analysis. We are going to combine the models of the weights in all possible ways (27 combinations) and compare the output of the weighted survival

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1				×	×	×	×	×	
CD4									
CD4 lag1									
Gender	×						×		
Age	×								
Base CD4	×								
Stage	×			×					
SL	×								
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1				×	×	×		×	
CD4									
CD4 lag1									
Gender	×								
Age	×								
Base CD4	×								
Stage									
SL	×								
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1				×	×	×	×	×	
CD4									
CD4 lag1									
Gender	×								
Age	×								
Base CD4	×								
Stage	×								
SL	×								
Total SL									

Table 5.13: Second best Models for the survival with AIC stepwise selection (left table, AIC=1863.2), AIC forward selection (middle table, AIC=1861.3) and the basic model (right table, AIC=1861.3).

analysis. As for the survival we are just going to use the basic model. The numbering of the combined models is given in Table 5.14.

SL	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3
cens	1	1	1	2	2	2	3	3	3	1	1	1	2	2	2	3	3	3	1	1	1	2	2	2	3	3	3
loss	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Model	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27

Table 5.14: Numbering of the models.

The coefficients for the second line treatment in the survival model is given in Figure 5.4. Note that the coefficient do not differ too much, the range is from -1.1507 to -1.0705 with a mean of -1.1208, which results in hazard ratios between 0.3164 and 0.3473 with mean 0.3260. Note that even the smallest negative value still is larger than the coefficient from the crude analysis.

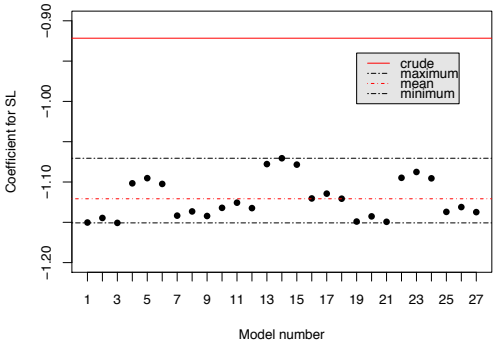

Figure 5.4: Plot of the coefficients for switch to second line treatment in the weighted analysis of survival.

Moreover we can see a structure in the plot of the coefficients.

First, we can see that there are always three points being very close. Within these three combination of models only the model for loss to follow-up changes. This shows us that it does not matter too much what model we choose for the loss to follow-up.

Second, there is again a structure within the first nine points (that repeats itself three times). We see that the middle triple is considerably higher than the one to the left and

to the right. Within these triples, the model for censoring changes, so we conclude that the model choice for the censoring is actually essential for the outcome.

Third, we can see that the three middle triples (points ten to 18) are in average higher than the part on the left and the right. This difference comes from the model choice for the switch of treatment. We can see that it does indeed influence the outcome, but not as much as the choice of the model for censoring does.

Impact of the Survival Models

We now want to see the impact of the choice of model for the survival. For this task we only use three combinations of the models for the weights, namely in our first combination there are the three models chosen with the AIC backwards criterion, in the second combination the three models with the AIC forward criterion and in the third combination the three basic models. We then accomplish the weighted analysis with the combinations of the models for the weights and the models for the survival (see Table 5.15).

SL/cens/lost	1/1/1	1/1/1	1/1/1	2/2/2	2/2/2	2/2/2	3/3/3	3/3/3	3/3/3
Survival	1	2	3	1	2	3	1	2	3
Model	1	2	3	4	5	6	7	8	9

Table 5.15: Numbering of the models.

The coefficients for the switch to second line treatment are given in Graph 5.5. The range is between -1.1675 and -1.0667 with mean -1.1227.

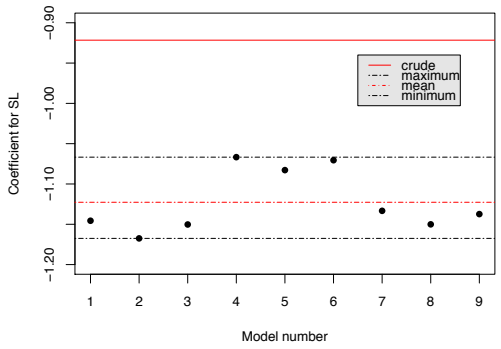

Figure 5.5: Plot of the coefficients for switch to second line treatment in the weighted analysis of survival.

As already in the first plot of the coefficients, there is an observable structure in these points. The middle three points are much higher than the rest. It is likely that the model for censoring is responsible for that circumstance, since we already saw that choosing the second model of censoring, results in a higher coefficient for the switch of treatment. Moreover we can see that the middle point of each triplet is lower than the other two. This is due to the second model for survival. However this difference is only small, so the main difference in the coefficients comes from the model choice for the weights.

5.5 Influential Data

In Figure 5.6 the weights of the three models are drawn on a log scale. Remember that we chose to use the stabilized weights because they are in a more narrow range around the value one. Note that the narrower these weights are to one, the smaller the variance of the estimates in the survival analysis. In this case however, there are some patient months which have a pretty large (or small) weights. This could be an indicator that these patients had a quite unlikely course of the CD4 count given their baseline values.

To understand this, note that to have an unlikely treatment, censoring of loss to follow-up value given the CD4 count, does not yet imply a big weight, because it is stabilized by multiplying it with the probability of treatment, censoring or loss to follow-up given the baseline values. Illustrated on the weights of the second line treatment, a large weight comes from the fact that

$$\mathrm{P}\left[T(t) = t(t) | \overline{C(t)} = \overline{c(t)}, \overline{B} = \bar{b}\right] \ll \mathrm{P}\left[T(t) = t(t) | \overline{B} = \bar{b}\right],$$

An example of that is a patient with a very low CD4 count at baseline, so the probability of switching is relatively high, but then experience a drastic increase of CD4 count so that switching treatment is pretty unlikely given the course of CD4 count over the time. When the opposite of that happens, the weight will be very close to zero.

It remains to say that a high or small weight could as well be due to the inaccurate specification of the models or unmeasured variables.

To determine the reason for large and small weights in this study, let us have a look at all the patients with weights above ten or below $\frac{1}{10}$.

Particularly there are six patients with weights above ten, two of them are lost to follow-up without switching treatment and the other four are switched to second line and censored at the end. There were no patient months with weights below $\frac{1}{10}$.

These six people all have a relatively high CD4 count when they are switched or lost to follow-up. Moreover, five out of these six have a low CD4 count at baseline, so this does support our guess of an unlikely course of CD4 count.

Since the weights of these six people are that high, they have a large impact in the weighted survival analysis. So in order to check the accuracy of the coefficient of the analysis from Chapter 6.1 we redo the analysis without these six patients. Note that the patients with small weights do barely influence the weighted analysis, but it is possible that they have an impact on the model selection for the model of the weights.

5.5.1 Analysis Revised

As for the crude analysis, the coefficient for second line is -0.9196 , which is almost identical to the coefficient for the whole data set.

The model selection results in the models in Table 5.16.

Note that only the second model for loss to follow-up differs from the model selection in the original analysis. All the other models turned out to be identical, even though the order of the drop of variables (add respectively) is not always the same.

The models for the survival are the same as in the model selection of the original data.

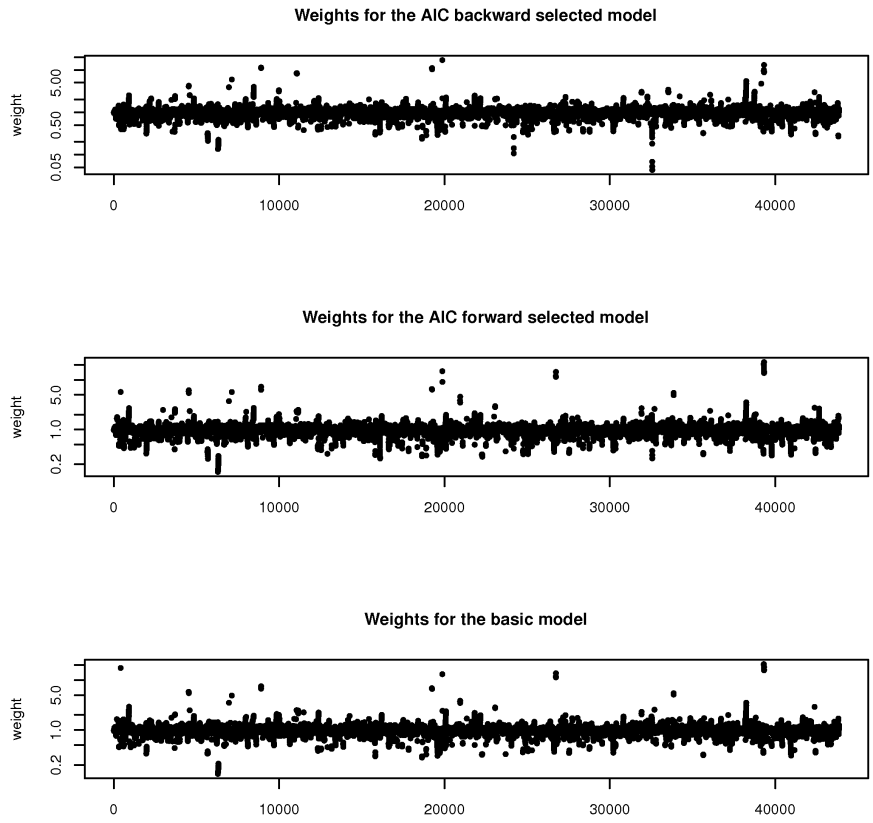

Figure 5.6: Plot of the weights on a log scale.

The black points in Graph 5.7 are the coefficients of the new analysis accomplished with the same model combinations as in Table 5.14. The gray point are the coefficients from the original analysis and the red line is for the coefficient of the crude analysis. As we can see, the coefficients are in general a little higher than in the original analysis, i.e. between -1.0404 and -1.1302. In this revised analysis it seems that the effect of switching is a little less beneficial than the coefficients of the original analysis implied, but still considerably more beneficial than the crude analysis implies.

As in the analysis of the original data, there is the same structure in this plot, i.e. the model choice for the loss to follow-up barely has an influence on the outcome, while the model choices of censoring and second line have a quite huge influence. Contrary to the original analysis, the model choice for the switch of treatment has a slightly greater influence on the outcome than the model choice of the censoring.

In Graph 5.8 the coefficients for the models from Table 5.15 are drawn. Again the values of the coefficients are in general higher than in the original analysis. The structure of the graph is the same as in the original analysis.

Graph 5.9 shows the weights of the new analysis without the six outliers. Again the outliers are mainly due to the weights of the switch of treatment and the loss to follow-up. There are again weights higher than ten for the second model. This comes from the fact that there is a different model for the loss to follow-up. The weights of the first model

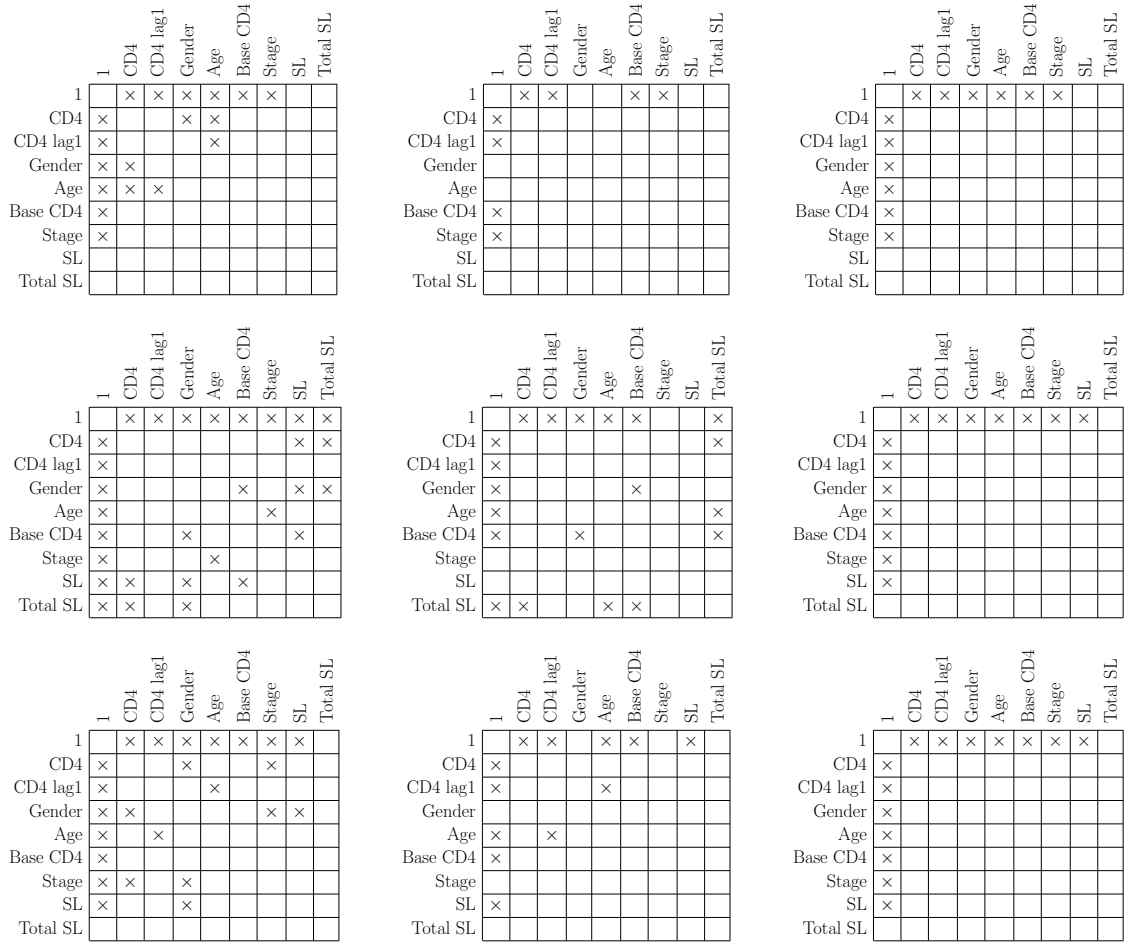

Table 5.16: Models for switch to second line (tables of first row), censoring (tables of second row) and loss to follow-up (tables of third row), with stepwise AIC backward criterion (tables of first column), AIC forward criterion (tables of second column) and the basic model (tables of third column.

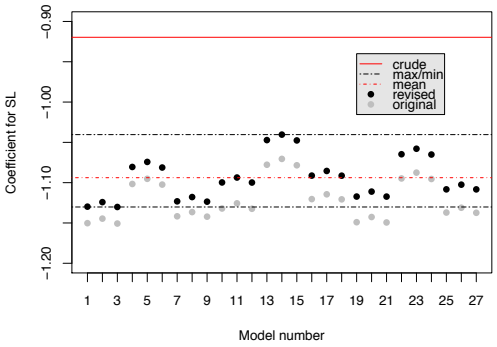

Figure 5.7: Plot of the coefficients for the switch to second line treatment.

combination are all below ten.

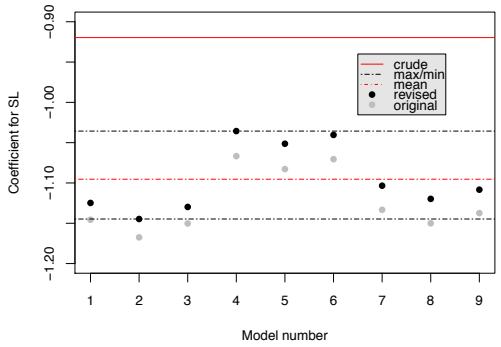

Figure 5.8: Plot of the coefficients for the switch to second line treatment.

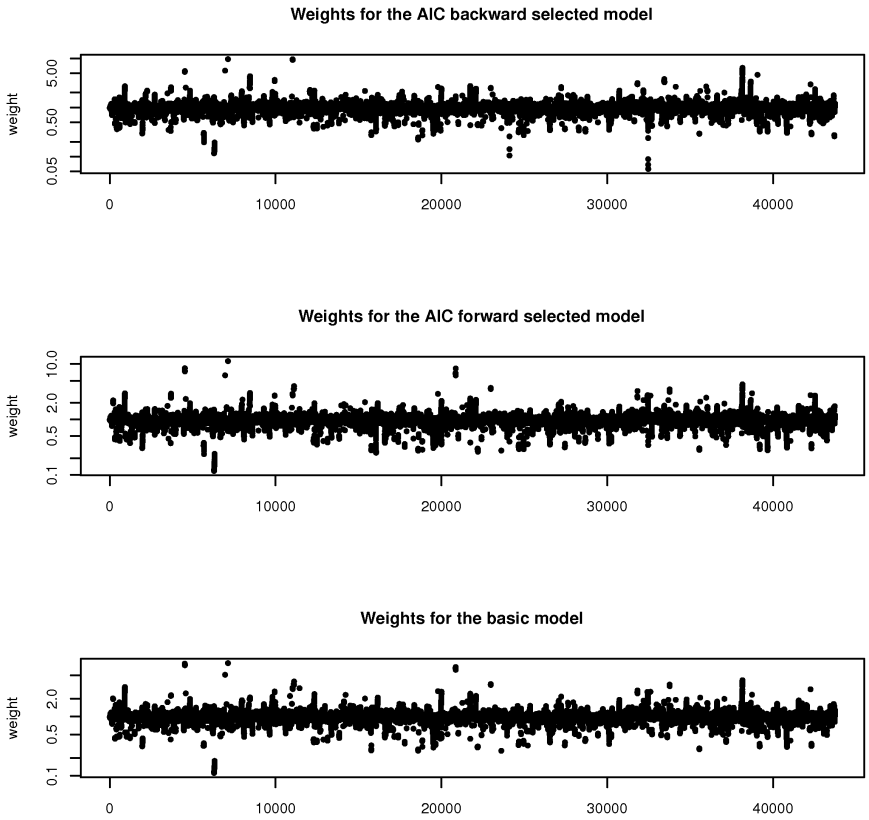

Figure 5.9: Plot of the weights.

Chapter 6

Conclusion

6.1 Results

The analysis showed that the treatment is indeed beneficial, more precisely the crude analysis implied that the switch to second line treatment results in a hazard ratio of 0.398 and the weighted analysis results in a hazard ratio of 0.321. Our initial idea was that the crude analysis underestimates the beneficial effect of the switch to second line treatment. That is so because we had reason to believe that mainly those patients who were in a bad health condition were switched to another treatment and hence the treated patients had a priori in average a lower live expectancy than the control group.

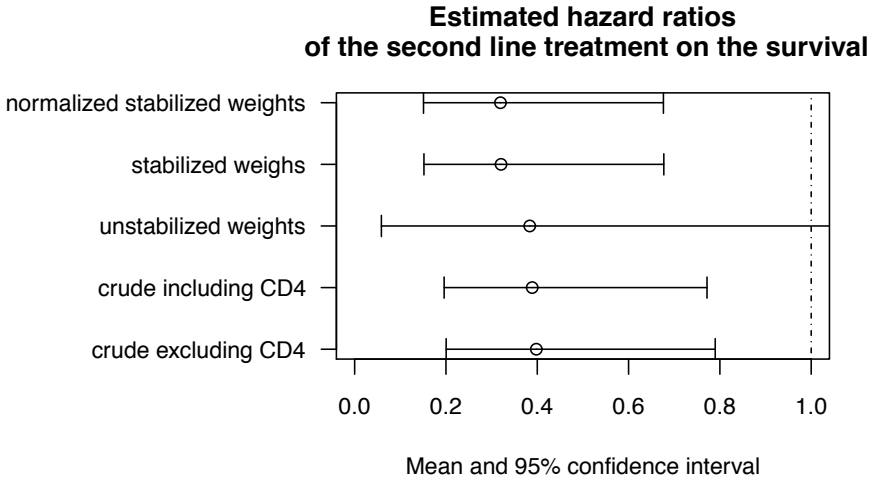

Figure 6.1: Plot of the estimated hazard ratios for the second line treatment with the different models.

In Figure 6.1 the results of the analysis are drawn together with the confidence intervals. The analysis with the unstabilized weights results in the largest confidence intervals, while the analysis with the stabilized and the normalized stabilized weights basically results in

the same hazard ratio and confidence intervals. The difference of the hazard ratio of the crude analysis and the weighted analysis is not as large as in other studies (like Hernan et al. (2000) and Sterne et al. (2005)) and by considering the relatively large confidence intervals this small difference is even less informative. This might be due to the fact that the CD4 count may not be the main influence on the decision to switch treatment. In fact, it seems like the decision to switch treatment is not really traceable with the available covariates.

6.2 Possible Caveats

The main problem in the analysis is the loss to follow-up. We have reason to believe that death influences the loss to follow-up. Remember from Graph 3.1 that an important assumption of marginal structural models is that the outcome does not influence the treatment. This assumption might be violated.

Moreover for the IPTW method to work, we assume that there are no unmeasured confounders for the treatment. But in fact, there might be such variables. The general health status and the medical site are the most explicit examples. An example of a health issue could be tuberculosis, which might have an impact on the decision to switch treatment.

Lastly, the proportional hazards assumption might be violated, i.e. β_1 in the definition of the hazard $h(t) = h_0(t) \cdot \exp(\beta_1 \cdot SL(t) + \overline{\beta} \cdot \overline{B})$ might be dependent on the time. In Chapter 5.2.5 we already checked if the total duration on second line or the time until the start of second line treatment has an (linear) influence on the outcome, and it turned out that the effect is negligible. However we could still check for a non linear dependence.

6.3 Possible Further Research

By using monthly data we have to insert a lot of data that is actually missing in our data set. One could chose more appropriate time intervals, such as quarterly data. This way one could avoid having to spuriously insert too much data.

Another further topic is sensitivity analysis for unmeasured confounders. Basically, one has to guess the direction of the influence of the unmeasured confounders on the outcome, based on the theoretical understanding of the matter and therewith simulate data of unmeasured confounders. By including this simulated variables in the analysis, one can see how much the results might be influenced by the unmeasured variables.

Finally, we want to indicate that there is another way of addressing marginal structural models, the so called G-Computation formula, which is another way of simulating data of a pseudo-population. The crucial step is that we have to estimate the probability of the observed confounder history given the treatment regime. This means we have to estimate the 'opposite' models than with the IPTW method.

Acknowledgement

I would like to thank Prof. Marloes Maathuis for the great mentoring and for the great choice of topic for my Masters Thesis.

Thanks to Matthias Egger, Thomas Gsponer and Olivia Keiser of the Institute for Social and Preventive Medicine in Bern for supplying me with the data, for the interesting discussions and the helpful inputs.

I would also like to thank Stephanie Werren and Weilian Shi for their help with Latex, R and the English language, and Deborah Flueck for her help with the Medical Background of HIV.

Finally, I would like to thank my parents, who always supported me in various ways throughout my years at ETH.

Bibliography

D'Agostino, R. B., M.-L. Lee, A. J. Belanger, L. A. Cupples, K. Anderson, and W. B. Kannel (1990). Relation of pooled logistic regression to time dependent cox regression analysis: The framingham heart study. *Statistics in Medicine 9*, 1501–1515.

Deutsche AIDS Gesellschaft e.V. and Osterreichische AIDS Gesellschaft (2010). Deutsch-Osterreichische Leitlinien zur antiretroviralen Therapie der HIV-1-Infektion. Website. Available online at http://www.rki.de/cln_160/nn_753398/DE/Content/InfAZ/H/HIVAIDS/Therapie/Leitlinien/D__A__antiretroviral__03__10,templateId=raw,property=publicationFile.pdf/D_A_antiretroviral_03_10.pdf.

Hartmann, D. M. (2008). HIV und AIDS, ein Leitfaden fuer Aerzte, Apotheker, Helfer und Betroffene. Website. Available online at http://www.hivleitfaden.de/cms/index.asp?hivleitfaden.

Hernan, M. A., B. Brumback, and J. M. Robins (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology 11*(5), 561–570.

Kleinbaum, D. G. and M. Klein (2005). *Survival Analysis, A Self-Learning Text*. Springer.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Robins, J. M., M. A. Hernan, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*(5), 550–560.

Sterne, J. A., M. A. Hernan, B. Ledergerber, K. Tilling, R. Weber, P. Sendi, M. Rickenbach, J. M. Robins, and M. Eger (2005). Long-term effectiveness of potent antiretroviral therapy in preventing aids and death: a prospective cohort study. *Lancet 366*, 378–384.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data, Extending the Cox Model*. Springer.

Thompson Jr., W. (1977). On the treatment of grouped observations in life studies. *Biometric 33*, 463–470.

Xiao, Y., M. Abrahamowicz, and E. E. Moodie (2010). Accuracy of conventional and marginal structural cox model estimators: A simulation study. *The International Journal of Biostatistics 6*(2), Article 13.

Appendix A

Pooled Logistic Regression

A.1 The Model

Thompson Jr. (1977) and D'Agostino et al. (1990) showed that one can use a so called pooled logistic model for survival data as an approximation of the cox model. Pooled means that we use different time intervals and in each interval we look at the number of objects at risk and the number of objects failing in this time interval. An example of that can be seen in Table A.1, where we assume we know the failure time for each object. The basic idea is then to fit a logistic model for each time interval.

Patient	Month of failure
1	3
2	2
3	3
4	2
5	5

Interval	Number at risk	Number of failures
$[0-1)$	5	0
$[1-2)$	5	2
$[2-3)$	3	2
$[3-4)$	1	0
$[4-5)$	1	1

Table A.1: Example for pooling the survival data of five objects.

We are going to comment on the case of time-independent covariates, but similarly to the Cox proportional hazards method one can also extend this theory to time-dependent covariates.

Remember that in logistic regression the dependent variable is a indicator variable F representing some kind of failure occurring in a pre-specified time period. If the value of F is one, then the patient failed some when in this time period. Note the difference to the survival analysis with the Cox model: The outcome is the failure time, while the outcome in logistic regression is an indicator of failure or no failure.

The logistic model is defined as:

$$\log\left(\frac{\mathrm{P}\left[F=1|\overline{X}\right]}{1-\mathrm{P}\left[F=1|\overline{X}\right]}\right)=\beta_0+\overline{\beta}^T\cdot\overline{X}$$

$$\iff \frac{\mathrm{P}\left[F=1|\overline{X}\right]}{1-\mathrm{P}\left[F=1|\overline{X}\right]}=\exp(\beta_0+\overline{\beta}^T\cdot\overline{X}).$$

Remember that the reason for choosing the logit link function is that the outcome is dichotomous and restricted to the interval $[0, 1]$.

The second equality is called the odds and represents the ratio of the probability of death and the probability of surviving. Odds can be interpreted the following way: If odds< 1 then the probability of death is smaller than the probability of surviving, i.e. the smaller the odds the better.

By using pooled logistic regression we fit a model for each time interval, so the outcome is not a single indicator variable for failure but rather a vector of indicator variables $F(t)$ for the failure at each interval $[t - \Delta t, t)$, where $t - \Delta t$ is the time of the last observation before t. If one assumes that the time has an influence on the survival probability, we might also include an appropriate function of the time in the linear part of the model. Note that $F(t)$ is only defined for those patients that did not fail before, hence the model actually fits $\mathrm{P}\left[F(t) = 1 | F(t-1) = 0\right]$, the conditional probability of failing in $[t - \Delta t, t)$ given that the failure did not occur earlier.

The model for the odds in the pooled logistic model is then:

$$\begin{aligned} o(t) :&= \frac{\mathrm{P}\left[F(t) = 1 | F(t-1) = 0, \overline{X}\right]}{1 - \mathrm{P}\left[F(t) = 1 | F(t) = 0, \overline{X}\right]} \\ &= \exp(\beta_0 + f(t) + \overline{\beta}^T \cdot \overline{X}) \\ &= \exp(\beta_0 + f(t)) \cdot \exp(\overline{\beta}^T \cdot \overline{X}) \\ &= o_0(t) \cdot \exp(\overline{\beta}^T \cdot \overline{X}), \end{aligned}$$

where $f(t)$ is some appropriate function of the time and similarly to the hazard function we call $o_0(t)$ the baseline odds.

Similarly to the Cox model the odds can be written as a product of a time dependent part and a part dependent on the covariates only. Therefore it makes sense to use the odds ratio as the measure of effect, which is then constant in time.

Assume that one of the covariates, say X_1 is again the treatment indicator. Let $\overline{X}_1 = \{X_{1,1}, X_{2,1}, \ldots X_{n,1}\}$ and $\overline{X}_2 = \{X_{1,2}, \ldots, X_{n,2}\}$ be the values of two cohorts respectively where the two are equal in all the covariates except the treatment. Let the first cohort be the treated group and the second the untreated, then

$$OR(\overline{X}_1, \overline{X}_2) := \frac{o_1(t)}{o_2(t)} = e^{\beta_1}.$$

The interpretation of the odds ratio is the same as the interpretation of the hazard ratio. If $OR(\overline{X}_1, \overline{X}_2) < 1$ (i.e. $\beta_1 < 0$) then the odds of the untreated is bigger than the odds of the treated.

This summarises to:

$$\beta_1 < 0 \Rightarrow \text{ treatment is successful.}$$

Remark. We still need to specify the function $f(t)$ of the time that we will include in the logistic regression. Note that in the Cox proportional hazards model the model for the baseline hazard was non-parametric. So we want to be as general as possible with the baseline odds as well. A good choice is to take the cubic splines of the time.

For more information on the splines, see for example Therneau and Grambsch (2000).

A.1.1 Implementation in R

In this case we have to convert the data a little further than we do for the Cox model. In order to receive survival curves for each time interval, we need to divide each persons observations into the same intervals. Otherwise, a person with some observation in $[0, 5)$ and another person with observation in $[3, 5)$ will count for the same interval.

Table A.2 shows an example of two patients with different follow-up times. In order to

patient	start	end	blood pressure	beta blocker	death
1	0	5	120	0	0
1	5	6	140	1	0
1	6	9	130	1	0
1	9	13	140	1	1
2	0	3	140	0	0
2	3	8	150	0	1
3	...				
⋮					

Table A.2: Data layout for the analysis of time-dependent covariates in survival analysis.

receive the same intervals for both patients, we have to convert the data into the data of Table A.3.

patient	start	end	blood pressure	beta blocker	death
1	0	3	120	0	0
1	3	5	120	0	0
1	5	6	140	1	0
1	6	8	130	1	0
1	8	9	130	1	0
1	9	13	140	1	1
2	0	3	140	0	0
2	3	5	150	0	0
2	5	6	150	0	0
2	6	8	150	0	1
3	...				
⋮					

Table A.3: Data layout for the analysis of time-dependent covariates in survival analysis for the conditional logistic model.

Even though this conversion is already sufficient for the analysis, we choose to divide each person into monthly periods. This does not improve the result of the analysis, but makes the interpretations of the coefficients easier as well as the application of Formula (4.1.0.1), since $(t_{k-1} - t_k) = 1$ for all $k \in \{1, \ldots, i\}$.

For this dataset we can now do the analysis with the following R commands:

```
formula <- death ~ blood pressure + beta blocker + ns(end,df)
model <- glm(formula, family=binomial())
```

where `ns(end,df)` is included in the package `splines` and calculates the B-Spline basis matrix of the time *end* of degree *df*.

Then in the summary of the model we have the coefficients of the dependent variables and also *df* variables for the splines of the time. There is no useful interpretation of the coefficients of the time, they are just needed to calculate the failure probability at each interval.

The probability of failure at each time point for a person is then:

```
pred <- predict(model, type="response")
```

By (4.1.0.1) the survivor curve can then be calculated as the product of all the past probabilities, i.e. $\mathrm{P}[T > t] = \prod_{s=1}^{t} \mathrm{P}[F(s) = 0] = \prod_{s=1}^{t}(1 - \mathrm{P}[F(s) = 1])$. This is easily done with the following function:

```
survivor <- cumprod(1-pred)
```

A.2 Analysis of the Data with the Pooled Logistic Model

We redo some parts of Chapter 5 with the pooled logistic model. We omit p-values and standard errors of the results in this chapter. For more detailed results see the outputs of the corresponding models in Appendix B. For the definition of the variables used in this chapter see Chapter 5.2.2.

A.2.1 Crude Analysis

We first want to do a crude analysis of the data to see the impact of the circumstance that the CD4 count is a confounder as well as a predictor of outcome. Clearly we have to include all the baseline covariates for the survivor function. The only question is if we want to include the CD4 count or not.

There is no 'right' answer to this question, since we are aware of the fact that the CD4 count is a confounder and an intermediate at the same time. Therefore we fit the two models:

$$\mathbf{E}[Y(t)|Y(t-1) = 0] = \frac{1}{1 + \exp(-(\alpha_0 + \alpha_1 \cdot T(t) + \overline{\alpha}^T \cdot \overline{B} + \overline{\beta}^T \cdot f(t)))}$$

$$\mathbf{E}[Y(t)|Y(t-1) = 0] = \frac{1}{1 + \exp(-(\alpha_0 + \alpha_1 \cdot C(t) + \alpha_2 \cdot T(t) + \overline{\alpha}^T \cdot \overline{B} + \overline{\beta}^T \cdot f(t)))}.$$

Note that since $Y(t)$ only has the outcomes 0 and 1, $\mathbf{E}[Y(t)] = \mathrm{P}[Y(t) = 1]$. As already mentioned, we take the cubic splines as the function of the time $f(t)$ (see Figure A.1). Note the decreasing character of $f(t)$. This implies that the odds of dying getting smaller as the time increase. Moreover we can see that the baseline odds are approximately linear, hence we would already have a pretty good fit if we would set $f(t) = log(t)$. The fact that $f(t)$ is always negative does not have a meaningful interpretation, since the baseline odds represent the hazard of a patient with CD4 count 0, which is obviously of no practical use.

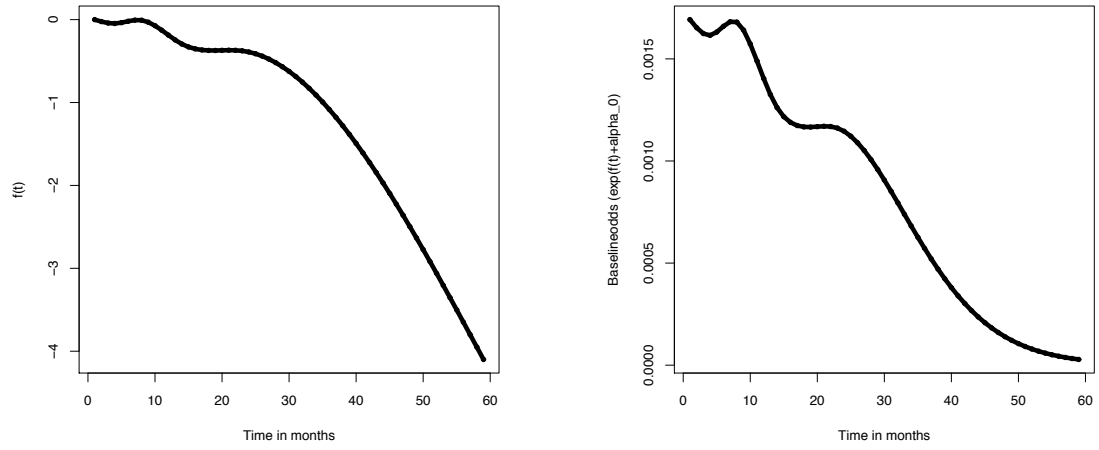

Figure A.1: Plot of the function f(t) and the plot of the baseline odds.

It turns out that the coefficient for the second line treatment does not differ to much in these two models. The coefficient is -0.925 and -0.947 for the two models. The interpretation of the coefficient is the following: A patient has a $e^{-0.925} = 0.3965$ times smaller odds in the next time period if he starts treatment rather than stays on first line treatment. The ratio from the other model only differ by a hundredth. The coefficient for the treatment if we choose the logarithm of the time instead of the spline function is -0.868.

A.2.2 IPTW Weighting with a Simple Model

As stated in Chapter 5.2.2 we have to estimates three weights, one for the switch to second line treatment, and for the censoring and one for the loss to follow-up.

Notation: In this chapter we are going to use the following notation:

$$Y(t) \sim X_1(t) + X_2(t) + \ldots + X_n(t)$$

for the survival model:

$$\mathbf{E}\left[Y(t)|Y(t-1)=0\right] = \frac{1}{1+\exp(-(\alpha_0 + \alpha_1 \cdot X_1(t) + \ldots + \alpha_n \cdot X_n(t)))},$$

to indicate that the indicator variable for failure at time t is dependent on the covariate values at that same time.

Weights for Second Line Treatment

We first want to derive $w_{i,T}(j)$ for all the patient i and months j. Note that in our study a patient stays on second line treatment forever when once started, hence

$$\mathrm{P}\left[T(j)=1|T(j-1)=1\right] = 1,$$

independent on the value of $\overline{T(j-2)}$ and $\overline{B}$. So we only need to fit a model for the part of the data up to the first initiation of second line treatment. One can easily see that under

this assumption our task is actually to fit a survival model where the start of second line treatment is the failure. So we get the following formulas for the weights:

- For j so that $t_i(j) = 0$ the weights simplify to:

$$w_{i,T}(j) = \frac{\mathrm{P}\left[T(j) = 0 | T(j-1) = 0, \overline{B} = \overline{b}_i\right]}{\mathrm{P}\left[T(j) = 0 | T(j-1) = 0, \overline{B} = \overline{b}_i, \overline{C(j)} = \overline{c_i(j)}\right]},$$

  with Equation (4.1.0.1).

- For the weights $w_{i,T}(j)$ with $t_i(j) = 1$, let $k = \min\{j | t_i(j) = 1\}$ be the first month of second line treatment. Then we can decompose the product:

$$\begin{aligned}
&\prod_{l=1}^{j} \mathrm{P}\left[T(l) = t(l) | \overline{T(l-1)} = \overline{t_i(l-1)}\right] \\
&= \Big(\prod_{l=1}^{k-1} \mathrm{P}\left[T(l) = 0 | T(l-1) = 0\right] \Big) \cdot \Big(\mathrm{P}\left[T(k) = 1 | T(k-1) = 0\right] \Big) \\
&\quad \cdot \Big(\prod_{l=k+1}^{j} \underbrace{\mathrm{P}\left[T(l) = 1 | T(l-1) = 1\right]}_{1} \Big) \\
&= \Big(\prod_{l=1}^{k-1} \mathrm{P}\left[T(l) = 0 | T(l-1) = 0\right] \Big) \cdot \Big(1 - \mathrm{P}\left[T(k) = 0 | T(k-1) = 0\right] \Big)
\end{aligned}$$

  Hence the weights $w_{i,T}(k), w_{i,T}(k+1), \ldots, w_{i,T}(t.end_i)$ are all equal and calculated as:

$$\begin{aligned}
w_i(j) = &\frac{\left(\prod_{l=1}^{k-1} \mathrm{P}\left[T(l) = 0 | T(l-1) = 0, \overline{b}_i\right] \right)}{\left(\prod_{l=1}^{k-1} \mathrm{P}\left[T(l) = 0 | T(l-1) = 0, \overline{b}_i, \overline{c_i(j)}\right] \right)} \\
&\times \frac{\left(1 - \mathrm{P}\left[T(k) = 0 | T(k-1) = 0, \overline{b}_i\right] \right)}{\left(1 - \mathrm{P}\left[T(k) = 0 | T(k-1) = 0, \overline{b}_i, \overline{c_i(j)}\right] \right)}
\end{aligned}$$

We can clearly not include the whole CD4 count history $\overline{C(j)}$ into the model, so we first going to do the analysis with only the CD4 count at current time $C(j)$, and then later include a lag variable.

So a the approach for the two models would be:

$$\begin{aligned}
T(t) &\sim \overline{B} \\
T(t) &\sim \overline{B} + C(t),
\end{aligned}$$

for the numerator and denominator of the stabilized weights respectively.

Weights for the Censoring and Lost to follow-up

Censoring and Lost to follow-up are both typical cases for survival analysis. They both happens once and there is no more information after the first failure. So the weights are

calculated the same way as the weights for the treatment, with the only difference that there is at most one weight to be calculated with the second equation.
For the model specification however, we have to take one more covariate into account. Obviously the treatment status can also influence the survivor function of censoring and lost to follow-up. The models for numerator and denominator for the weights of censoring are then:

$$T'(t) \sim T(t) + \overline{B}$$
$$T'(t) \sim T(t) + \overline{B} + C(t),$$

and the models for the loss to follow-up are exactly the same:

$$T''(t) \sim T(t) + \overline{B}$$
$$T''(t) \sim T(t) + \overline{B} + C(t).$$

Analysis with the Weights

By weighting the population with the product of the three weights we assure that the treatment is no longer confounded by the CD4 count. Hence we can apply the pooled logistic model to the weighted data without having the problem described in the crude analysis. The model we fit in the pseudo-population is then:

$$Y(t) \sim T(t) + \overline{B}$$

The resulting coefficient is -1.21938, implying that a patient has 0.2954 times the odds in the next time period if he starts second line treatment rather than stays on first line treatment.
Compared to the ratio of the crude estimation (0.3965), the weighted analysis implies that the switch to second line treatment is slightly more beneficial than the crude analysis implies.

A.2.3 Interpretation

So far we only interpreted the causal coefficient of the treatment, but clearly there are more coefficients to interpret, namely the intercept and the coefficients for the baseline covariates.
The baseline covariates are:

- Gender: 1 for male and 0 for female
- Age: Age group (<30, 30-39, $\geq$ 40) at the beginning of the study
- CD4: CD4 group (<50, 50-99, 100-199, $\geq$ 200) at the date of immunological failure
- Stage: Indicator if disease is in an advanced stage

The output of the weighted pooled logistic regression is shown in Table A.4.

Then the interpretations of the coefficients of the IPTW model are:

Model	Intercept	SL	Gender 1	Age		CD4			Stage 0
				<30	$\geq$ 40	<50	50-99	100-199	
crude	-6.385	-0.925	-0.107	-0.024	-0.108	1.967	1.485	0.939	-0.329
weighted	-6.382	-1.219	-0.134	-0.028	-0.111	2.015	1.485	0.943	-0.338

Table A.4: Coefficients for the simplest model.

- If we compare the odds of a male and a female with the same treatment, age group, CD4 baseline group and Stage, then the male has $e^{-0.134} = 0.875$ times the odds of the woman, i.e. the survival is slightly better for men.
- The two negative coefficients for the age groups imply that the patients younger than 30 and older than 40 both have a better survival than the ones aged 30 to 39, i.e. they have $e^{-0.028} = 0.972$ and $e^{-0.111} = 0.895$ times the odds of the patients aged between 30 and 39. Hence the patients aged between 30 and 39 have the worst survival, and the ones above 40 have the best survival expectancy.
- As for the CD4 count the coefficients imply that the higher the value the better the survival. The three groups have a ratio of $e^{0.943} = 2.568$, $e^{1.485} = 4.415$ and $e^{2.015} = 7.501$ compared to the group with a CD4 count above 200.
- The coefficient for the stage implies that a patient who is not in advanced stage at the beginning of the study has $e^{-0.338} = 0.713$ times the odds of a patient who was in advanced stage when all other covariates are the same. Hence the expected survival is better for patients not in advanced stage.

Note that the effect of CD4 and Stage on the survival is comprehensible while the effect of the age group and gender is not explicable with logical consideration, but since the effect is comparable small, we believe that this output is actually reasonable.

A.3 Sensitivity to Model Selection

In general we need to fit four models for the analysis of the data set. A model for the switch to second line treatment, a model for censoring, a model for loss to follow-up and a model for the survival.

The model from Chapter A.2.2 is only the most simplest one, but other terms could be important to include in the analysis too, in particular interaction terms and some covariates describing the CD4 history.

As an additional covariate we include the $C.lag1(t)$, which is the CD4 count at month $t-1$. As already mentioned we actually need to control for an appropriate function of the CD4 history, not only the current value, so including the CD4 count one month before is the most obvious thing to do.

Moreover we include the total duration of second line treatment. The reason for that is that we believe that not only the switch to second line, but also the total duration of second line treatment could have an influence on the censoring and the loss to follow-up.

Notation: In the following we are going to illustrate the models with matrices as in Table A.5.

	1	X_1	X_2
1	×	×	×
X_1	×		×
X_2	×	×	

Table A.5: Example of a Model.

The first row and column represent the main effects. In this example we have that both variables X_1 and X_2 are in the model. The cross in the first square of the first row stands for the intercept and all the other squares for the interaction terms. If the example above were the model for Y, then together with the notation from Chapter A.2.2, the Table A.5 reads as:

$$Y \sim X_1 + X_2 + X_1 \cdot X_2$$

A.3.1 Models for the weights

In the following we are going to select three models for the treatment, censoring and lost to follow-up. The first is chosen with the AIC step backward criterion, the second model with the AIC step forward criterion and the third is the full model without any interaction terms. We then do the IPTW analysis with all possible combinations of these models.

In these three cases we are going to define a maximal model including all the reasonable covariates and interaction terms, and a minimal model including the covariates we believe have to be in the model even if the criteria are not fulfilled. The maximal models are given in Table A.6.

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×	×	×	×	×		
CD4	×			×	×		×		
CD4 lag1	×			×	×		×		
Gender	×	×	×		×	×	×		
Age	×	×	×	×		×	×		
Base CD4	×			×	×		×		
Stage	×	×	×	×	×	×			
SL									
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×	×	×	×	×	×	×
CD4	×			×	×		×	×	×
CD4 lag1	×			×	×		×	×	×
Gender	×	×	×		×	×	×	×	×
Age	×	×	×	×		×	×	×	×
Base CD4	×			×	×		×	×	×
Stage	×	×	×	×	×	×		×	×
SL	×	×	×	×	×	×	×		
Total SL	×	×	×	×	×	×	×		

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×	×	×	×	×	×	×
CD4	×			×	×		×	×	×
CD4 lag1	×			×	×		×	×	×
Gender	×	×	×		×	×	×	×	×
Age	×	×	×	×		×	×	×	×
Base CD4	×			×	×		×	×	×
Stage	×	×	×	×	×			×	×
SL	×	×	×	×	×	×	×		
Total SL	×	×	×	×	×	×	×		

Table A.6: Maximal model for second line treatment (AIC=3822.5), the censoring (AIC=18172) and the lost to follow-up (AIC=1745.7).

The reason for not including the interaction terms of CD4 count, CD4 lag1 and baseline CD4 is that they are difficult to interpret.

Note that in all the three cases we need a model for denominator and numerator simultaneously, i.e. for the numerator we are just going to use the same model but reduced, i.e. excluding $C(t)$ and $C.lag1(t)$. It turns out that the baseline CD4 count is highly significant in all the three reduced models, but not at all in the models for the denominator. Therefore the baseline CD4 count is one covariate we decide to include in the minimal model. Moreover the CD4 count at current time is the essential covariate that we want to control for, hence it is also included in the minimal model. Thus the minimal model is the same in the three cases and includes the baseline CD4 count and the CD4 count at current time.

Models for the Second Line Treatment

Table A.7 gives the three models. Note that the CD4 count and the CD4 count at the previous month both have a negative coefficient. This means that the higher the CD4 count the smaller the odds of switching as we expected it.

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×	×	×	×	×		
CD4	×			×	×				
CD4 lag1	×				×				
Gender	×	×							
Age	×	×	×						
Base CD4	×								
Stage	×								
SL									
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×			×	×		
CD4	×								
CD4 lag1	×								
Gender									
Age									
Base CD4	×								
Stage	×								
SL									
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×	×	×	×	×		
CD4	×								
CD4 lag1	×								
Gender	×								
Age	×								
Base CD4	×								
Stage	×								
SL									
Total SL									

Table A.7: Best Models for second line treatment with AIC backward selection (left table, AIC=3797),AIC forward selection (middle table, AIC=3807.4) and the basic model (right table, AIC=3811).

Notes on unmeasured variables with influence: We do not really know on what the decision to switch to second line treatment really depends on. According to the WHO criteria on the CD4 count all the patient should have been switched at the beginning of the study, but since this is not so, the switching seems somewhat arbitrary. An obvious confounder is the general health condition of a patient. Moreover, probable unmeasured variables of influence could be personal bias of the doctors, site of the medical center, availability of the second line treatment, wealth of the patient and so on.

Models for the Censoring

Table A.8 gives the three models.

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×	×	×	×	×	×	×
CD4	×							×	×
CD4 lag1	×								
Gender	×					×		×	×
Age	×						×		
Base CD4	×			×				×	×
Stage	×				×				
SL	×	×		×		×			
Total SL	×	×		×		×			

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×	×	×	×			×
CD4	×								×
CD4 lag1	×								
Gender	×					×			
Age	×								
Base CD4	×			×					×
Stage									
SL									
Total SL	×	×				×			

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×	×	×	×	×	×	
CD4	×								
CD4 lag1	×								
Gender	×								
Age	×								
Base CD4	×								
Stage	×								
SL	×								
Total SL									

Table A.8: Best Models for the censoring with AIC backward selection(left table, AIC=18139), AIC forward selection (middle table, AIC=18149) and the basic model (right table, AIC=18175).

Notes on unmeasured variables with influence: Censoring basically correlates with better life expectancy, hence we could expect that only personal features of the patient influence the censoring. The most important covariates are already included in the baseline variables, but one more important variable could be the location of the patient.

Models for the Loss to Follow-up

Table A.9 gives the three models. Note that the coefficient for the CD4 count is negative. This implies that the lower the CD4 count the more likely someone is lost to follow-up.

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×	×	×	×	×	×	
CD4	×			×			×		
CD4 lag1	×				×				
Gender	×	×					×	×	
Age	×		×					×	
Base CD4	×								
Stage	×	×		×					
SL	×			×	×				
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×			×	×		×	
CD4	×				×				
CD4 lag1									
Gender									
Age	×	×						×	
Base CD4	×								
Stage									
SL	×				×				
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×	×	×	×	×	×	×	×	
CD4	×								
CD4 lag1	×								
Gender	×								
Age	×								
Base CD4	×								
Stage	×								
SL	×								
Total SL									

Table A.9: Best Models for the lost to follow-up with AIC backward selection (left table, AIC=1701.3), AIC forward selection (middle table, AIC=1706.9) and the basic model (right table, AIC=1716.1).

Notes on unmeasured variables with influence: For the loss to follow-up we believe that there is a lot of influence coming from unmeasured variables. First there are features of the patient itself influencing the lost to follow-up. Examples are: wealth, general health, location, relocation, confidence in the medic in charge, and so on. Second, the medical center can account for some more unmeasured confounder, features like preciseness with the collection of the data and how hard they try to contact a patient not coming back will clearly influence the amount of patients lost to follow-up.

Finally death itself is a main confounder. The fact that a low CD4 count increases the probability of being lost to follow-up is an indicator that sicker people are more likely to be lost to follow-up, and sicker people are also more likely to be dead.

A.3.2 Models for the Survival

In epidemiologic studies, the gender and age is always an important covariate, so it has to be in the minimal model. Also the switch to second line treatment has to be in the model because it is the goal variable. The maximal model is shown in Table A.10. Clearly the time dependent CD4 count variables are not allowed to be in the model. Moreover we do not want to include the total duration of second line and any interaction terms with second line, because it makes it hard to interpret the effect of the switch.

The AIC forward and backward selection both result in the basic model. So in order to have three different models, we are going to use the second best model from the AIC backward and forward selections. The last variable that is dropped in the backwards selection is the interaction term of advanced stage and gender, so our fist model includes this interaction term together with all the main effects of the basic model.

The last variable that is added in the forward selection is the advanced stage, so the second model includes all the main effects of the basic models except the advanced stage. Note that the second model is the same model that we would obtain by the BIC stepwise selection. The three models for survival are given in table A.11.

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×			×	×			×	
CD4									
CD4 lag1									
Gender	×								
Age	×								
Base CD4									
Stage									
SL	×								
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×			×	×	×	×	×	
CD4									
CD4 lag1									
Gender	×				×	×	×		
Age	×			×		×	×		
Base CD4	×			×	×		×		
Stage	×			×	×	×			
SL	×								
Total SL									

Table A.10: Minimal (left table) and maximal (right table) model of the weighted analysis for the survival.

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×			×	×	×	×	×	
CD4									
CD4 lag1									
Gender	×						×		
Age	×								
Base CD4	×								
Stage	×			×					
SL	×								
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×			×	×	×		×	
CD4									
CD4 lag1									
Gender	×								
Age	×								
Base CD4	×								
Stage									
SL	×								
Total SL									

	1	CD4	CD4 lag1	Gender	Age	Base CD4	Stage	SL	Total SL
1	×			×	×	×	×	×	
CD4									
CD4 lag1									
Gender	×								
Age	×								
Base CD4	×								
Stage	×								
SL	×								
Total SL									

Table A.11: Second best Models for the survival with AIC stepwise selection (left table, AIC=1863.2), AIC forward selection (middle table, AIC=1861.3) and the basic model (right table, AIC=1861.3).

A.3.3 Results

Impact of the Treatment Models

We first want to see the impact of the model choice for the weights on the output of the weighted survival analysis. We are going to combine the models of the weights in all possible ways (27 combinations) and compare the output of the weighted survival analysis. As for the survival we are just going to use the basic model. The numbering of the combined models is given in Table A.12.

SL	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3
cens	1	1	1	2	2	2	3	3	3	1	1	1	2	2	2	3	3	3	1	1	1	2	2	2	3	3	3
loss	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Model	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27

Table A.12: Numbering of the models.

The coefficients for the second line treatment in the survival model is given in Figure A.2. Note that the coefficient do not differ too much, the range is from -1.1499 to -1.0878, which results in odds ratios between 0.3167 and 0.3370 and even the smallest negative value is considerably larger than the coefficient from the crude analysis.

Moreover we can see a structure in the plot of the coefficients.

First, we can see that there are always three points being very close. Within these three combination of models only the model for loss to follow-up changes. This shows us that

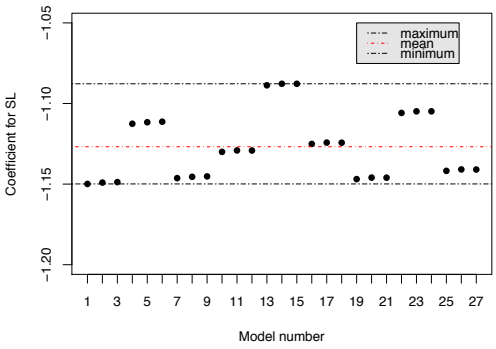

Figure A.2: Plot of the coefficients for switch to second line treatment in the weighted analysis of survival.

it does not matter too much what model we choose for the loss to follow-up.

Second, there is again a structure within the first nine points (that repeats itself three times). We see that the middle triple is considerably higher than the one to the left and to the right. Within these triples, the model for censoring changes, so we conclude that the model choice for the censoring is actually essential for the outcome.

Third, we can see that the three middle triples (points ten to 18) are in average higher than the part on the left and the right. This difference comes from the model choice for the switch of treatment. We can see that it does indeed influence the outcome, but not as much as the choice of the model for censoring does.

Impact of the Survival Models

We now want to see the impact of the choice of model for the survival. For this task we only use three combinations of the models for the weights, namely in our first combination there are the three models chosen with the AIC backwards criterion, in the second combination the three models with the AIC forward criterion and in the third combination the three basic models. We then accomplish the weighted analysis with the combinations of the models for the weights and the models for the survival (see Table A.13).

SL/cens/lost	1/1/1	1/1/1	1/1/1	2/2/2	2/2/2	2/2/2	3/3/3	3/3/3	3/3/3
Survival	1	2	3	1	2	3	1	2	3
Model	1	2	3	4	5	6	7	8	9

Table A.13: Numbering of the models.

The coefficients for the switch to second line treatment are given in Graph A.3. The range is between -1.1658 and -1.0842.

As already in the former plot of the coefficients, there is an observable structure in these points. The middle three points are much higher than the rest. It is likely that the model for censoring is responsible for that circumstance, since we already saw that choosing the second model of censoring, results in a higher coefficient for the switch of treatment. Moreover we can see that the middle point of each triplet is lower than the other two.

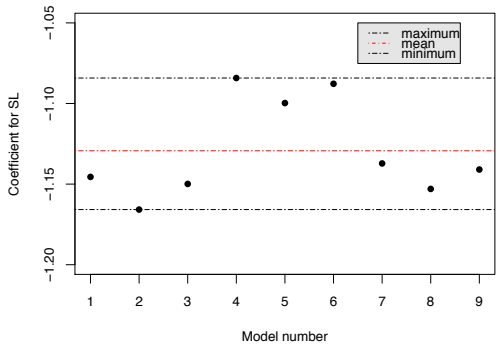

Figure A.3: Plot of the coefficients for switch to second line treatment in the weighted analysis of survival.

This is due to the second model for survival. However this difference is only small, so the main difference in the coefficients comes from the model choice for the weights.

A.4 Model Checking

A.4.1 The Weights

In Figure A.4 the weights of the three models are drawn on a log scale. Remember that we chose to use the stabilized weights because they are in a more narrow range around the value one. Note that the narrower these weights are to one, the smaller the variance of the estimates in the survival analysis. In this case however, there are some patient months which have a pretty large (or small) weight. This could be an indicator that these patients had a quite unlikely course of the CD4 count given their baseline values.

To understand this, note that to have an unlikely treatment, censoring of loss to follow-up value given the CD4 count, does not yet imply a big weight, because it is stabilized by multiplying it with the probability of treatment, censoring or loss to follow-up given the baseline values. So a large weight comes from the fact that

$$\mathrm{P}\left[T_t = t_t | \overline{CD}_t = \overline{cd}_t, B = b\right] \ll \mathrm{P}\left[T_t = t_t | B = b\right],$$

where T_t stands for switch of treatment, censoring or loss to follow-up at time t. An example of that is a patient with a very low CD4 count at baseline, so the probability of switching is relatively high, but then experience a drastic increase of CD4 count so that switching treatment is pretty unlikely given the course of CD4 count over the time. When the opposite of that happens, the weight will be very close to zero.

It remains to say that a high or small weight could as well be due to the inaccurate specification of the models or unmeasured variables.

To determine the reason for large and small weights in this study, let us have a look at all the patients with weights above ten or below $\frac{1}{10}$.

Particularly there are six patients with weights above ten, two of them are lost to follow-up without switching treatment and the other four are switched to second line and censored at the end. There were no patient months with weights below $\frac{1}{10}$.

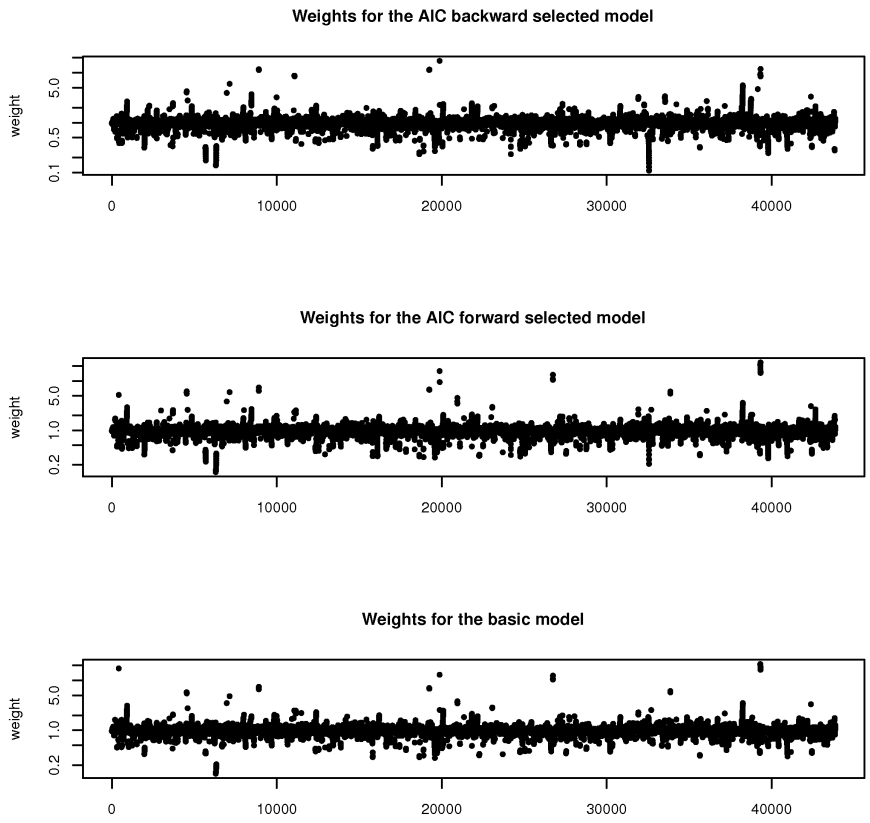

Figure A.4: Plot of the weights.

These six people all have a relatively high CD4 count when they are switched or lost to follow-up. Moreover, five out of these six have a low CD4 count at baseline, so this does support our guess of an unlikely course of CD4 count.

Since the weights of these six people are that high, they have a large impact in the weighted survival analysis. So in order to check the accuracy of the coefficient of the analysis from Chapter 6.1 we redo the analysis without these six patients. Note that the patients with small weights do barely influence the weighted analysis, but it is possible that they have an impact on the model selection for the model of the weights.

Analysis Revised

As for the crude analysis, the coefficient for second line is -0.923, which is almost identical to the coefficient for the whole data set.

The model selection results in the models in Table 5.16.

Note that only the first model for censoring and the second model for loss to follow-up differ. All the other models turned out to be identical to the ones in the original analysis, even though the order of the drop of variables (add respectively) is not the same.

The black points in Graph A.5 are the coefficients of the new analysis accomplished with the same model combinations as in Table 5.14. The grey points are the coefficients from

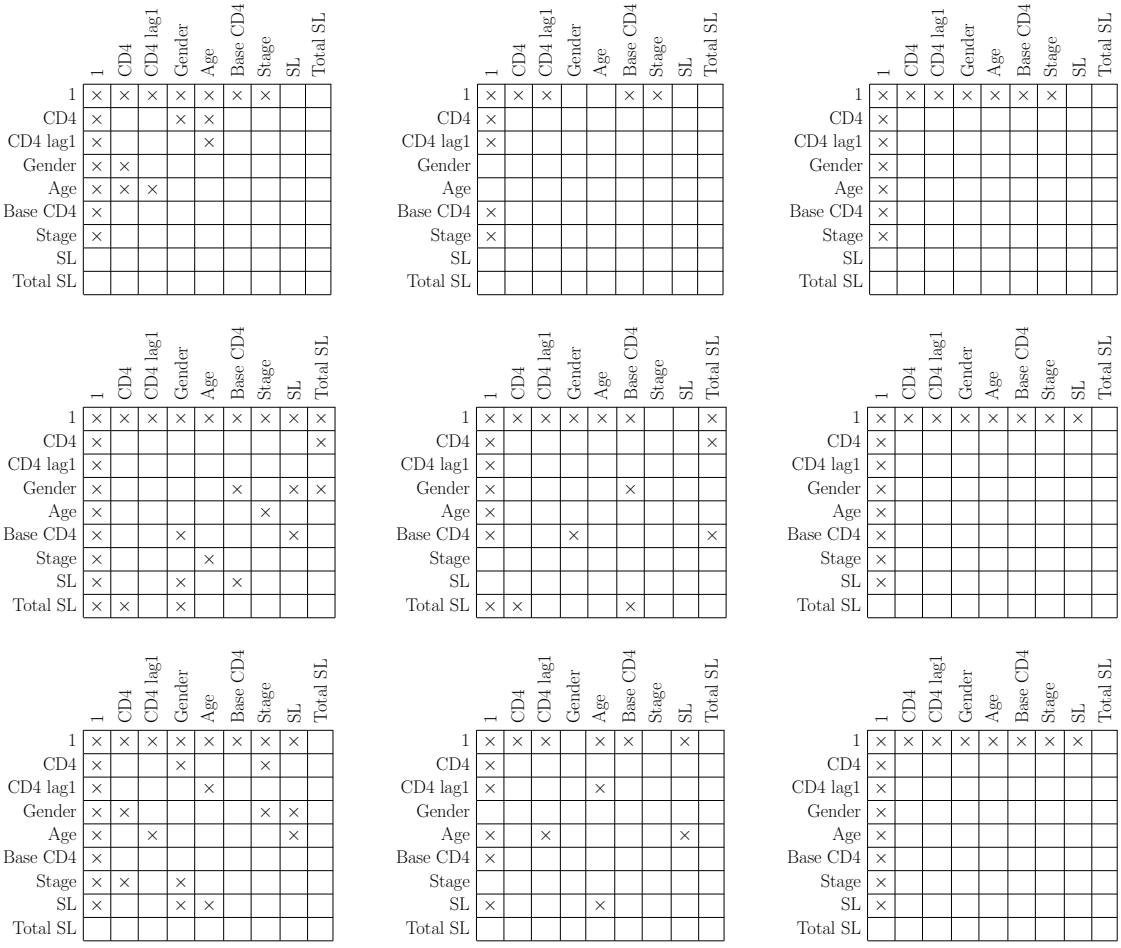

Table A.14: Models for switch to second line (tables of first row), censoring (tables of second row) and loss to follow-up (tables of third row), with stepwise AIC backward criterion (tables of first column), AIC forward criterion (tables of second column) and the basic model (tables of third column.

the original analysis and the red line is for the coefficient of the crude analysis (in the original data set). As we can see, the coefficients are in general a little higher than in the original analysis, i.e. between -1.1230 and -1.0613. With this revised analysis it seems that the effect of switching is a little less beneficial than the coefficients of the original analysis implied, but still considerably more beneficial than the crude analysis implies.

As in the analysis of the original data, there is the same structure in this plot, i.e. the model choice for the loss to follow-up barely has an influence on the outcome, while the model choices of censoring and second line have a quite huge influence. Contrary to the original analysis, the model choice for the switch of treatment has a slightly greater influence on the outcome than the model choice of the censoring.

In Graph A.6 the coefficients for the models from Table A.13 are drawn. Again the values of the coefficients are in general higher than in the original analysis. The structure of the graph is the same as in the original analysis.

Graph A.7 shows the weights of the new analysis without the six outliers. Again the outliers are mainly due to the weights of the switch of treatment and the loss to follow-up.

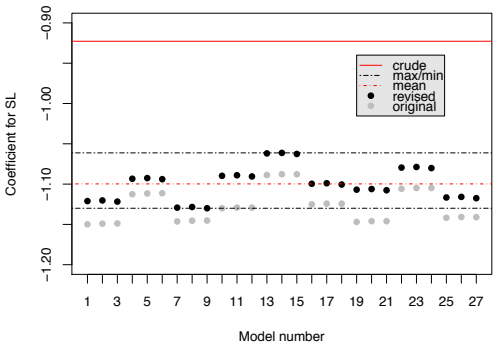

Figure A.5: Plot of the coefficients for the switch to second line treatment.

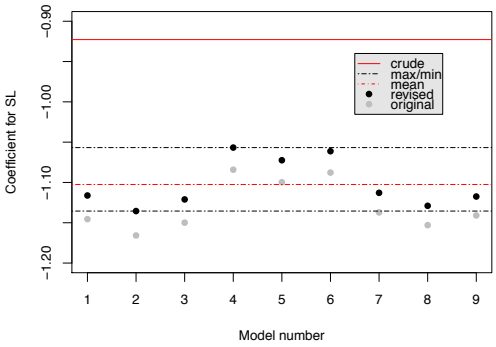

Figure A.6: Plot of the coefficients for the switch to second line treatment.

There are again weights higher than ten for the second model. This comes from the fact that there is a different model for the loss to follow-up. The weights of the first model combination are all below ten.

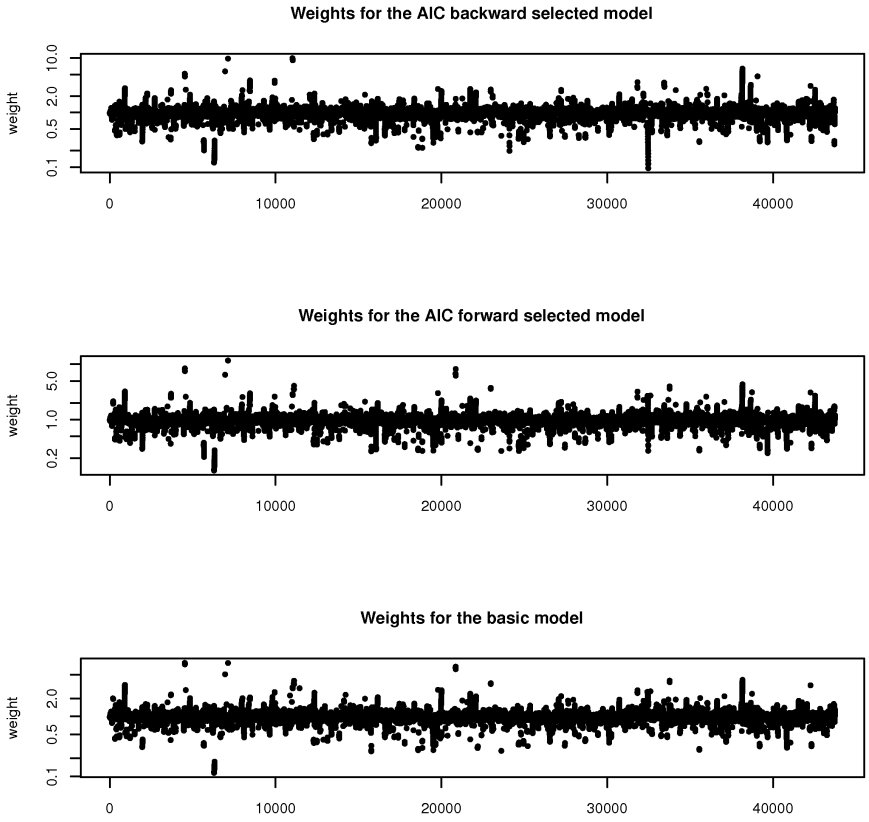

Figure A.7: Plot of the weights.

Appendix B

Outputs

B.1 Outputs for Chapter 5

B.1.1 Outputs for Chapter 5.2.1

Output for $S \sim T(t) + \overline{B}$

```
                      coef exp(coef) se(coef)      z Pr(>|z|)
sl_yn             -0.92158   0.39789  0.34983 -2.634  0.00843 **
gender1           -0.10811   0.89753  0.17163 -0.630  0.52875
bas_age_gple30    -0.02303   0.97723  0.19092 -0.121  0.90397
bas_age_gpgteq40  -0.10617   0.89927  0.38717 -0.274  0.78391
bas_cd4_gp50_99    1.48186   4.40112  0.30323  4.887 1.02e-06 ***
bas_cd4_gp100_199  0.93730   2.55309  0.31426  2.983  0.00286 **
bas_cd4_gple50     1.96295   7.12032  0.31330  6.265 3.72e-10 ***
adv_stage0        -0.32961   0.71921  0.21216 -1.554  0.12029
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
sl_yn                0.3979     2.5133    0.2004    0.7898
gender1              0.8975     1.1142    0.6411    1.2564
bas_age_gple30       0.9772     1.0233    0.6722    1.4207
bas_age_gpgteq40     0.8993     1.1120    0.4210    1.9207
bas_cd4_gp50_99      4.4011     0.2272    2.4292    7.9739
bas_cd4_gp100_199    2.5531     0.3917    1.3790    4.7268
bas_cd4_gple50       7.1203     0.1404    3.8532   13.1577
adv_stage0           0.7192     1.3904    0.4745    1.0900

Rsquare= 0.001   (max possible= 0.046 )
Likelihood ratio test= 62.02  on 8 df,   p=1.872e-10
Wald test            = 54.59  on 8 df,   p=5.3e-09
Score (logrank) test = 62.61  on 8 df,   p=1.431e-10
```

Output for $S \sim C(t) + T(t) + \overline{B}$

```
                        coef  exp(coef)   se(coef)      z Pr(>|z|)
cd4                -0.0032601  0.9967452  0.0009384 -3.474 0.000512 ***
sl_yn              -0.9444565  0.3888909  0.3496929 -2.701 0.006917 **
gender1            -0.0477843  0.9533394  0.1719105 -0.278 0.781043
bas_age_gple30      0.0047880  1.0047995  0.1908945  0.025 0.979990
bas_age_gpgteq40   -0.0481674  0.9529743  0.3873233 -0.124 0.901031
bas_cd4_gp50_99     0.8799380  2.4107503  0.3425335  2.569 0.010202 *
bas_cd4_gp100_199   0.5501467  1.7335073  0.3296991  1.669 0.095190 .
bas_cd4_gple50      1.2921932  3.6407629  0.3641158  3.549 0.000387 ***
adv_stage0         -0.3060672  0.7363371  0.2121931 -1.442 0.149190
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

                   exp(coef) exp(-coef) lower .95 upper .95
cd4                   0.9967     1.0033    0.9949    0.9986
sl_yn                 0.3889     2.5714    0.1960    0.7718
gender1               0.9533     1.0489    0.6806    1.3353
bas_age_gple30        1.0048     0.9952    0.6912    1.4607
bas_age_gpgteq40      0.9530     1.0493    0.4461    2.0360
bas_cd4_gp50_99       2.4108     0.4148    1.2319    4.7176
bas_cd4_gp100_199     1.7335     0.5769    0.9084    3.3080
bas_cd4_gple50        3.6408     0.2747    1.7834    7.4324
adv_stage0            0.7363     1.3581    0.4858    1.1161

Rsquare= 0.002   (max possible= 0.046 )
Likelihood ratio test= 77.14  on 9 df,   p=5.964e-13
Wald test            = 66.08  on 9 df,   p=8.896e-11
Score (logrank) test = 73.34  on 9 df,   p=3.358e-12
```

Output for $SL \sim C(t) + \overline{B}$

```
                        coef  exp(coef)   se(coef)      z Pr(>|z|)
cd4                -0.0085469  0.9914896  0.0009777 -8.742   <2e-16 ***
bas_cd4_gp50_99    -0.2201093  0.8024311  0.2312202 -0.952    0.341
bas_cd4_gp100_199   0.0213350  1.0215642  0.1983419  0.108    0.914
bas_cd4_gple50      0.3001317  1.3500366  0.2519705  1.191    0.234
bas_age_gple30     -0.0734924  0.9291432  0.1164264 -0.631    0.52require(splines)
require(geneplotter)

#-------------------------------------------------------------------------
data.dir <- "/u/peterk/Desktop/Data_Set"
ds4 <- dget(paste(data.dir,"data_total.dmp",sep="/"))

L <- length(ds4$patient) #Number of rows
l <- ds4$patient[L]      #Number of patients8
bas_age_gpgteq40   -0.2564923  0.7737610  0.2544055 -1.008    0.313
```

```
adv_stage0         0.1820331  1.1996538  0.1157756  1.572    0.116
gender1            0.1230538  1.1309453  0.1053384  1.168    0.243
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                 exp(coef) exp(-coef) lower .95 upper .95
cd4                 0.9915     1.0086    0.9896    0.9934
bas_cd4_gp50_99     0.8024     1.2462    0.5100    1.2625
bas_cd4_gp100_199   1.0216     0.9789    0.6925    1.5069
bas_cd4_gple50      1.3500     0.7407    0.8239    2.2122
bas_age_gple30      0.9291     1.0763    0.7396    1.1673
bas_age_gpgteq40    0.7738     1.2924    0.4700    1.2740
adv_stage0          1.1997     0.8336    0.9561    1.5052
gender1             1.1309     0.8842    0.9200    1.3903

Rsquare= 0.007   (max possible= 0.133 )
Likelihood ratio test= 265.1  on 8 df,   p=0
Wald test            = 204.5  on 8 df,   p=0
Score (logrank) test = 229.3  on 8 df,   p=0
```

B.1.2 Outputs for Chapter 5.2.2

Output for $SL \sim \overline{B}$

```
                     coef exp(coef) se(coef)      z Pr(>|z|)
bas_cd4_gp50_99    1.23644   3.44333  0.17949  6.889 5.64e-12 ***
bas_cd4_gp100_199  0.88768   2.42948  0.18080  4.910 9.12e-07 ***
bas_cd4_gple50     1.98440   7.27467  0.18005 11.021  < 2e-16 ***
bas_age_gple30    -0.09202   0.91209  0.11634 -0.791    0.429
bas_age_gpgteq40  -0.32797   0.72038  0.25435 -1.289    0.197
adv_stage0         0.13140   1.14043  0.11565  1.136    0.256
gender1            0.04541   1.04645  0.10539  0.431    0.667
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                 exp(coef) exp(-coef) lower .95 upper .95
bas_cd4_gp50_99     3.4433     0.2904    2.4221     4.895
bas_cd4_gp100_199   2.4295     0.4116    1.7046     3.463
bas_cd4_gple50      7.2747     0.1375    5.1115    10.353
bas_age_gple30      0.9121     1.0964    0.7261     1.146
bas_age_gpgteq40    0.7204     1.3882    0.4376     1.186
adv_stage0          1.1404     0.8769    0.9091     1.431
gender1             1.0465     0.9556    0.8512     1.287

Rsquare= 0.004   (max possible= 0.133 )
Likelihood ratio test= 148.9  on 7 df,   p=0
Wald test            = 140.5  on 7 df,   p=0
Score (logrank) test = 166.2  on 7 df,   p=0
```

Output for $SL \sim \overline{B} + C(t) + C.lag1(t)$

```
                       coef exp(coef)  se(coef)      z Pr(>|z|)
cd4                -0.005255  0.994759  0.001762 -2.983  0.00286 **
cd4lag1            -0.003901  0.996107  0.001827 -2.135  0.03276 *
bas_cd4_gp50_99    -0.346720  0.707004  0.238667 -1.453  0.14630
bas_cd4_gp100_199  -0.065799  0.936319  0.202200 -0.325  0.74487
bas_cd4_gple50      0.133974  1.143363  0.264541  0.506  0.61255
bas_age_gple30     -0.071859  0.930662  0.116453 -0.617  0.53719
bas_age_gpgteq40   -0.256193  0.773993  0.254401 -1.007  0.31391
adv_stage0          0.184095  1.202130  0.115784  1.590  0.11184
gender1             0.125645  1.133880  0.105352  1.193  0.23302
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                   exp(coef) exp(-coef) lower .95 upper .95
cd4                   0.9948     1.0053    0.9913    0.9982
cd4lag1               0.9961     1.0039    0.9925    0.9997
bas_cd4_gp50_99       0.7070     1.4144    0.4429    1.1287
bas_cd4_gp100_199     0.9363     1.0680    0.6300    1.3917
bas_cd4_gple50        1.1434     0.8746    0.6808    1.9203
bas_age_gple30        0.9307     1.0745    0.7407    1.1693
bas_age_gpgteq40      0.7740     1.2920    0.4701    1.2743
adv_stage0            1.2021     0.8319    0.9581    1.5084
gender1               1.1339     0.8819    0.9223    1.3939

Rsquare= 0.007   (max possible= 0.133 )
Likelihood ratio test= 269.7  on 9 df,   p=0
Wald test            = 205.7  on 9 df,   p=0
Score (logrank) test = 230.9  on 9 df,   p=0
```

Output for $Cens \sim T(t) + \overline{B}$

```
                        coef exp(coef) se(coef)      z Pr(>|z|)
sl_yn               0.35095   1.42041  0.05901  5.947 2.73e-09 ***
bas_cd4_gp50_99    -0.39766   0.67189  0.05674 -7.009 2.41e-12 ***
bas_cd4_gp100_199  -0.12700   0.88073  0.05121 -2.480 0.013134 *
bas_cd4_gple50     -0.25826   0.77239  0.06952 -3.715 0.000203 ***
bas_age_gple30      0.07175   1.07439  0.04686  1.531 0.125729
bas_age_gpgteq40    0.20809   1.23133  0.09132  2.279 0.022683 *
adv_stage0          0.03841   1.03916  0.04620  0.831 0.405743
gender1            -0.09947   0.90532  0.04210 -2.363 0.018132 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                   exp(coef) exp(-coef) lower .95 upper .95
sl_yn                 1.4204     0.7040    1.2653    1.5946
bas_cd4_gp50_99       0.6719     1.4883    0.6012    0.7509
bas_cd4_gp100_199     0.8807     1.1354    0.7966    0.9737
```

```
bas_cd4_gple50       0.7724     1.2947    0.6740    0.8852
bas_age_gple30       1.0744     0.9308    0.9801    1.1777
bas_age_gpgteq40     1.2313     0.8121    1.0295    1.4727
adv_stage0           1.0392     0.9623    0.9492    1.1376
gender1              0.9053     1.1046    0.8336    0.9832

Rsquare= 0.002   (max possible= 0.526 )
Likelihood ratio test= 86.43  on 8 df,   p=2.442e-15
Wald test            = 86.96  on 8 df,   p=1.887e-15
Score (logrank) test = 87.44  on 8 df,   p=1.554e-15
```

Output for $Cens \sim T(t) + \overline{B} + C(t) + C.lag1(t)$

```
                         coef  exp(coef)   se(coef)      z Pr(>|z|)
sl_yn                0.3379091  1.4020131  0.0592009  5.708 1.14e-08 ***
cd4                 -0.0012064  0.9987943  0.0003203 -3.766 0.000166 ***
cd4lag1              0.0007207  1.0007210  0.0003093  2.330 0.019802 *
bas_cd4_gp50_99     -0.4835120  0.6166140  0.0628427 -7.694 1.42e-14 ***
bas_cd4_gp100_199   -0.1878497  0.8287392  0.0547374 -3.432 0.000600 ***
bas_cd4_gple50      -0.3440681  0.7088807  0.0749780 -4.589 4.46e-06 ***
bas_age_gple30       0.0763320  1.0793209  0.0468752  1.628 0.103438
bas_age_gpgteq40     0.2035675  1.2257679  0.0913447  2.229 0.025843 *
adv_stage0           0.0454828  1.0465330  0.0462426  0.984 0.325327
gender1             -0.0833026  0.9200727  0.0422911 -1.970 0.048868 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                   exp(coef) exp(-coef) lower .95 upper .95
sl_yn                 1.4020     0.7133    1.2484    1.5745
cd4                   0.9988     1.0012    0.9982    0.9994
cd4lag1               1.0007     0.9993    1.0001    1.0013
bas_cd4_gp50_99       0.6166     1.6218    0.5452    0.6974
bas_cd4_gp100_199     0.8287     1.2067    0.7444    0.9226
bas_cd4_gple50        0.7089     1.4107    0.6120    0.8211
bas_age_gple30        1.0793     0.9265    0.9846    1.1832
bas_age_gpgteq40      1.2258     0.8158    1.0248    1.4661
adv_stage0            1.0465     0.9555    0.9559    1.1458
gender1               0.9201     1.0869    0.8469    0.9996

Rsquare= 0.002   (max possible= 0.526 )
Likelihood ratio test= 106.1  on 10 df,   p=0
Wald test            = 106.9  on 10 df,   p=0
Score (logrank) test = 106.6  on 10 df,   p=0
```

Output for $Loss \sim T(t) + \overline{B}$

```
                       coef exp(coef) se(coef)      z Pr(>|z|)
sl_yn              -1.39873   0.24691  0.46013 -3.040  0.00237 **
```

```
bas_cd4_gp50_99    0.71630   2.04685  0.25376  2.823  0.00476 **
bas_cd4_gp100_199  0.45246   1.57217  0.25569  1.770  0.07680 .
bas_cd4_gple50     0.83700   2.30944  0.29505  2.837  0.00456 **
bas_age_gple30     0.36618   1.44222  0.21299  1.719  0.08556 .
bas_age_gpgteq40  -0.14689   0.86339  0.48373 -0.304  0.76139
adv_stage0        -0.04684   0.95424  0.20295 -0.231  0.81749
gender1           -0.15449   0.85685  0.17949 -0.861  0.38939
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
sl_yn                0.2469     4.0501    0.1002    0.6084
bas_cd4_gp50_99      2.0469     0.4886    1.2448    3.3658
bas_cd4_gp100_199    1.5722     0.6361    0.9525    2.5950
bas_cd4_gple50       2.3094     0.4330    1.2953    4.1177
bas_age_gple30       1.4422     0.6934    0.9500    2.1894
bas_age_gpgteq40     0.8634     1.1582    0.3345    2.2282
adv_stage0           0.9542     1.0479    0.6411    1.4204
gender1              0.8568     1.1671    0.6027    1.2181

Rsquare= 0.001   (max possible= 0.041 )
Likelihood ratio test= 26.72  on 8 df,   p=0.0007909
Wald test            = 22.62  on 8 df,   p=0.003882
Score (logrank) test = 24.02  on 8 df,   p=0.002276
```

Output for $Loss \sim T(t) + \overline{B} + C(t) + C.lag1(t)$

```
                       coef exp(coef)  se(coef)      z Pr(>|z|)
sl_yn             -1.450481  0.234458  0.459788 -3.155  0.00161 **
cd4               -0.002443  0.997560  0.002066 -1.183  0.23682
cd4lag1           -0.002447  0.997556  0.002141 -1.143  0.25300
bas_cd4_gp50_99   -0.161929  0.850502  0.296092 -0.547  0.58446
bas_cd4_gp100_199 -0.115035  0.891335  0.272800 -0.422  0.67326
bas_cd4_gple50    -0.154801  0.856586  0.348491 -0.444  0.65690
bas_age_gple30     0.403470  1.497010  0.213224  1.892  0.05846 .
bas_age_gpgteq40  -0.079906  0.923203  0.483603 -0.165  0.86876
adv_stage0         0.000208  1.000208  0.203174  0.001  0.99918
gender1           -0.047722  0.953399  0.179836 -0.265  0.79073
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
sl_yn                0.2345     4.2652   0.09521    0.5773
cd4                  0.9976     1.0024   0.99353    1.0016
cd4lag1              0.9976     1.0025   0.99338    1.0018
bas_cd4_gp50_99      0.8505     1.1758   0.47604    1.5195
bas_cd4_gp100_199    0.8913     1.1219   0.52220    1.5214
bas_cd4_gple50       0.8566     1.1674   0.43265    1.6959
bas_age_gple30       1.4970     0.6680   0.98566    2.2736
```

```
bas_age_gpgteq40     0.9232     1.0832   0.35781    2.3820
adv_stage0           1.0002     0.9998   0.67166    1.4895
gender1              0.9534     1.0489   0.67019    1.3563

Rsquare= 0.001   (max possible= 0.041 )
Likelihood ratio test= 60.04  on 10 df,   p=3.563e-09
Wald test            = 47.13  on 10 df,   p=8.954e-07
Score (logrank) test = 47.37  on 10 df,   p=8.102e-07
```

Output for $S \sim T(t) + \overline{B}$ (weighted)

```
                      coef exp(coef) se(coef)      z Pr(>|z|)
sl_yn             -1.13736   0.32066  0.38140 -2.982  0.00286 **
gender1           -0.12061   0.88638  0.17008 -0.709  0.47825
bas_age_gple30    -0.05175   0.94956  0.18790 -0.275  0.78299
bas_age_gpgteq40  -0.12495   0.88254  0.38501 -0.325  0.74554
bas_cd4_gp50_99    1.49380   4.45397  0.30116  4.960 7.04e-07 ***
bas_cd4_gp100_199  0.93492   2.54700  0.31285  2.988  0.00280 **
bas_cd4_gple50     2.01403   7.49348  0.30993  6.498 8.12e-11 ***
adv_stage0        -0.34715   0.70670  0.21184 -1.639  0.10127
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
sl_yn                0.3207     3.1185    0.1518    0.6772
gender1              0.8864     1.1282    0.6351    1.2371
bas_age_gple30       0.9496     1.0531    0.6570    1.3724
bas_age_gpgteq40     0.8825     1.1331    0.4150    1.8770
bas_cd4_gp50_99      4.4540     0.2245    2.4683    8.0370
bas_cd4_gp100_199    2.5470     0.3926    1.3795    4.7025
bas_cd4_gple50       7.4935     0.1334    4.0820   13.7560
adv_stage0           0.7067     1.4150    0.4666    1.0704

Rsquare= 0.002   (max possible= 0.047 )
Likelihood ratio test= 69.99  on 8 df,   p=4.928e-12
Wald test            = 61.14  on 8 df,   p=2.782e-10
Score (logrank) test = 70.19  on 8 df,   p=4.504e-12
```

B.1.3 Outputs for Chapter 5.2.4

Output for the interaction of gender and second line

```
                      coef exp(coef) se(coef)      z Pr(>|z|)
gender1           -0.14568   0.86443  0.17693 -0.823  0.41029
bas_age_gple30    -0.02621   0.97413  0.19094 -0.137  0.89082
bas_age_gpgteq40  -0.11130   0.89467  0.38719 -0.287  0.77376
bas_cd4_gp50_99    1.48135   4.39888  0.30321  4.886 1.03e-06 ***
bas_cd4_gp100_199  0.93582   2.54931  0.31427  2.978  0.00290 **
```

```
bas_cd4_gple50     1.96808   7.15692  0.31311  6.286 3.27e-10 ***
adv_stage0        -0.32537   0.72226  0.21222 -1.533  0.12523
sl_yn:gender0     -1.21530   0.29662  0.51943 -2.340  0.01930 *
sl_yn:gender1     -0.61502   0.54063  0.46648 -1.318  0.18736
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
gender1              0.8644     1.1568    0.6111     1.223
bas_age_gple30       0.9741     1.0266    0.6700     1.416
bas_age_gpgteq40     0.8947     1.1177    0.4189     1.911
bas_cd4_gp50_99      4.3989     0.2273    2.4280     7.970
bas_cd4_gp100_199    2.5493     0.3923    1.3769     4.720
bas_cd4_gple50       7.1569     0.1397    3.8744    13.220
adv_stage0           0.7223     1.3845    0.4765     1.095
sl_yn:gender0        0.2966     3.3713    0.1072     0.821
sl_yn:gender1        0.5406     1.8497    0.2167     1.349

Rsquare= 0.001   (max possible= 0.046 )
Likelihood ratio test= 62.77  on 9 df,   p=3.91e-10
Wald test            = 55.42  on 9 df,   p=1.011e-08
Score (logrank) test = 63.73  on 9 df,   p=2.547e-10
```

Output for the interaction of base age and second line

```
                           coef exp(coef) se(coef)      z Pr(>|z|)
bas_age_gple30         -0.08696   0.91671  0.19482 -0.446  0.65534
bas_age_gpgteq40       -0.22759   0.79645  0.41034 -0.555  0.57913
bas_cd4_gp50_99         1.48969   4.43571  0.30317  4.914 8.94e-07 ***
bas_cd4_gp100_199       0.93562   2.54878  0.31428  2.977  0.00291 **
bas_cd4_gple50          1.96435   7.13025  0.31345  6.267 3.68e-10 ***
adv_stage0             -0.32583   0.72193  0.21220 -1.535  0.12467
gender1                -0.11192   0.89412  0.17146 -0.653  0.51393
sl_yn:bas_age_gp30_39  -2.03103   0.13120  1.01469 -2.002  0.04532 *
sl_yn:bas_age_gple30   -0.68928   0.50194  0.39748 -1.734  0.08289 .
sl_yn:bas_age_gpgteq40 -0.10336   0.90180  1.07281 -0.096  0.92324
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                       exp(coef) exp(-coef) lower .95 upper .95
bas_age_gple30            0.9167     1.0909   0.62575    1.3430
bas_age_gpgteq40          0.7964     1.2556   0.35635    1.7801
bas_cd4_gp50_99           4.4357     0.2254   2.44852    8.0357
bas_cd4_gp100_199         2.5488     0.3923   1.37663    4.7190
bas_cd4_gple50            7.1303     0.1402   3.85742   13.1799
adv_stage0                0.7219     1.3852   0.47628    1.0943
gender1                   0.8941     1.1184   0.63892    1.2512
sl_yn:bas_age_gp30_39     0.1312     7.6220   0.01796    0.9586
sl_yn:bas_age_gple30      0.5019     1.9923   0.23031    1.0939
```

```
sl_yn:bas_age_gpgteq40    0.9018     1.1089   0.11014    7.3839

Rsquare= 0.001   (max possible= 0.046 )
Likelihood ratio test= 64.57  on 10 df,   p=4.907e-10
Wald test            = 55.51  on 10 df,   p=2.533e-08
Score (logrank) test = 64.29  on 10 df,   p=5.55e-10
```

Output for the interaction of base CD4 count and second line

```
                              coef  exp(coef)   se(coef)      z Pr(>|z|)
bas_cd4_gp50_99           1.483e+00  4.405e+00  3.043e-01  4.872 1.10e-06 ***
bas_cd4_gp100_199         8.994e-01  2.458e+00  3.167e-01  2.840  0.00451 **
bas_cd4_gple50            1.932e+00  6.903e+00  3.172e-01  6.091 1.12e-09 ***
adv_stage0               -3.286e-01  7.199e-01  2.122e-01 -1.549  0.12142
gender1                  -1.069e-01  8.986e-01  1.717e-01 -0.623  0.53333
bas_age_gple30           -2.655e-02  9.738e-01  1.910e-01 -0.139  0.88941
bas_age_gpgteq40         -1.100e-01  8.958e-01  3.872e-01 -0.284  0.77633
sl_yn:bas_cd4_gpgteq200 -1.423e+01  6.627e-07  1.529e+03 -0.009  0.99257
sl_yn:bas_cd4_gp50_99   -1.313e+00  2.690e-01  7.231e-01 -1.816  0.06937 .
sl_yn:bas_cd4_gp100_199 -4.799e-01  6.189e-01  7.281e-01 -0.659  0.50985
sl_yn:bas_cd4_gple50    -8.111e-01  4.444e-01  4.799e-01 -1.690  0.09098 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                         exp(coef) exp(-coef) lower .95 upper .95
bas_cd4_gp50_99          4.405e+00  2.270e-01    2.4260     7.997
bas_cd4_gp100_199        2.458e+00  4.068e-01    1.3215     4.573
bas_cd4_gple50           6.903e+00  1.449e-01    3.7071    12.853
adv_stage0               7.199e-01  1.389e+00    0.4750     1.091
gender1                  8.986e-01  1.113e+00    0.6418     1.258
bas_age_gple30           9.738e-01  1.027e+00    0.6698     1.416
bas_age_gpgteq40         8.958e-01  1.116e+00    0.4194     1.914
sl_yn:bas_cd4_gpgteq200 6.627e-07  1.509e+06    0.0000       Inf
sl_yn:bas_cd4_gp50_99    2.690e-01  3.718e+00    0.0652     1.110
sl_yn:bas_cd4_gp100_199 6.189e-01  1.616e+00    0.1485     2.579
sl_yn:bas_cd4_gple50     4.444e-01  2.250e+00    0.1735     1.138

Rsquare= 0.001   (max possible= 0.046 )
Likelihood ratio test= 63.23  on 11 df,   p=2.312e-09
Wald test            = 53.14  on 11 df,   p=1.696e-07
Score (logrank) test = 65.26  on 11 df,   p=9.627e-10
```

Output for the interaction of advanced stage and second line

```
                     coef exp(coef) se(coef)      z Pr(>|z|)
adv_stage0        -0.43672   0.64615  0.22665 -1.927  0.05400 .
gender1           -0.10218   0.90287  0.17168 -0.595  0.55173
bas_age_gple30    -0.01989   0.98030  0.19086 -0.104  0.91698
```

```
bas_age_gpgteq40  -0.10116   0.90379  0.38724 -0.261  0.79391
bas_cd4_gp50_99    1.47913   4.38911  0.30330  4.877 1.08e-06 ***
bas_cd4_gp100_199  0.93881   2.55692  0.31425  2.987  0.00281 **
bas_cd4_gple50     1.96511   7.13569  0.31300  6.278 3.42e-10 ***
sl_yn:adv_stage1  -1.30281   0.27177  0.46137 -2.824  0.00475 **
sl_yn:adv_stage0  -0.04839   0.95277  0.54423 -0.089  0.92915
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
adv_stage0           0.6462     1.5476    0.4144    1.0075
gender1              0.9029     1.1076    0.6449    1.2640
bas_age_gple30       0.9803     1.0201    0.6744    1.4250
bas_age_gpgteq40     0.9038     1.1065    0.4231    1.9306
bas_cd4_gp50_99      4.3891     0.2278    2.4222    7.9532
bas_cd4_gp100_199    2.5569     0.3911    1.3811    4.7337
bas_cd4_gple50       7.1357     0.1401    3.8638   13.1783
sl_yn:adv_stage1     0.2718     3.6796    0.1100    0.6713
sl_yn:adv_stage0     0.9528     1.0496    0.3279    2.7684

Rsquare= 0.001   (max possible= 0.046 )
Likelihood ratio test= 64.92  on 9 df,   p=1.499e-10
Wald test            = 56.99  on 9 df,   p=5.087e-09
Score (logrank) test = 65.36  on 9 df,   p=1.226e-10
```

Output for the interaction of gender and second line (weighted)

```
                     coef exp(coef) se(coef)      z Pr(>|z|)
gender1           -0.14245   0.86723  0.17442 -0.817  0.41411
bas_age_gple30    -0.05199   0.94934  0.18788 -0.277  0.78199
bas_age_gpgteq40  -0.12835   0.87954  0.38504 -0.333  0.73887
bas_cd4_gp50_99    1.49426   4.45603  0.30114  4.962 6.98e-07 ***
bas_cd4_gp100_199  0.93636   2.55068  0.31285  2.993  0.00276 **
bas_cd4_gple50     2.01844   7.52655  0.30987  6.514 7.33e-11 ***
adv_stage0        -0.34460   0.70851  0.21189 -1.626  0.10389
sl_yn:gender0     -1.33662   0.26273  0.54526 -2.451  0.01423 *
sl_yn:gender1     -0.91183   0.40179  0.52788 -1.727  0.08411 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
gender1              0.8672     1.1531   0.61613     1.221
bas_age_gple30       0.9493     1.0534   0.65690     1.372
bas_age_gpgteq40     0.8795     1.1370   0.41353     1.871
bas_cd4_gp50_99      4.4560     0.2244   2.46954     8.040
bas_cd4_gp100_199    2.5507     0.3921   1.38151     4.709
bas_cd4_gple50       7.5265     0.1329   4.10045    13.815
adv_stage0           0.7085     1.4114   0.46771     1.073
sl_yn:gender0        0.2627     3.8062   0.09024     0.765
```

```
sl_yn:gender1        0.4018     2.4889   0.14278     1.131

Rsquare= 0.002   (max possible= 0.047 )
Likelihood ratio test= 70.31  on 9 df,   p=1.324e-11
Wald test            = 61.67  on 9 df,   p=6.392e-10
Score (logrank) test = 71.24  on 9 df,   p=8.712e-12
```

Output for the interaction of base age and second line (weighted)

```
                          coef exp(coef) se(coef)      z Pr(>|z|)
bas_age_gple30         -0.1035    0.9017   0.1913 -0.541  0.58858
bas_age_gpgteq40       -0.2443    0.7832   0.4075 -0.600  0.54882
bas_cd4_gp50_99         1.5047    4.5028   0.3011  4.997 5.83e-07 ***
bas_cd4_gp100_199       0.9342    2.5451   0.3129  2.986  0.00283 **
bas_cd4_gple50          2.0133    7.4877   0.3101  6.492 8.48e-11 ***
adv_stage0             -0.3465    0.7072   0.2119 -1.635  0.10197
gender1                -0.1224    0.8848   0.1700 -0.720  0.47161
sl_yn:bas_age_gp30_39  -2.1198    0.1201   0.9951 -2.130  0.03315 *
sl_yn:bas_age_gple30   -0.9080    0.4033   0.4473 -2.030  0.04236 *
sl_yn:bas_age_gpgteq40 -0.1591    0.8529   1.0830 -0.147  0.88320
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                       exp(coef) exp(-coef) lower .95 upper .95
bas_age_gple30            0.9017     1.1090   0.61981    1.3118
bas_age_gpgteq40          0.7832     1.2767   0.35240    1.7409
bas_cd4_gp50_99           4.5028     0.2221   2.49552    8.1247
bas_cd4_gp100_199         2.5451     0.3929   1.37846    4.6990
bas_cd4_gple50            7.4877     0.1336   4.07729   13.7508
adv_stage0                0.7072     1.4141   0.46685    1.0712
gender1                   0.8848     1.1302   0.63409    1.2347
sl_yn:bas_age_gp30_39     0.1201     8.3292   0.01708    0.8441
sl_yn:bas_age_gple30      0.4033     2.4793   0.16786    0.9692
sl_yn:bas_age_gpgteq40    0.8529     1.1725   0.10210    7.1248

Rsquare= 0.002   (max possible= 0.047 )
Likelihood ratio test= 72.25  on 10 df,   p=1.631e-11
Wald test            = 61.77  on 10 df,   p=1.670e-09
Score (logrank) test = 71.81  on 10 df,   p=1.980e-11
```

Output for the interaction of base CD4 count and second line (weighted)

```
                             coef  exp(coef)   se(coef)      z Pr(>|z|)
bas_cd4_gp50_99          1.497e+00  4.466e+00  3.019e-01  4.957 7.15e-07 ***
bas_cd4_gp100_199        8.971e-01  2.453e+00  3.150e-01  2.848   0.0044 **
bas_cd4_gple50           1.977e+00  7.219e+00  3.133e-01  6.310 2.80e-10 ***
adv_stage0              -3.443e-01  7.087e-01  2.118e-01 -1.625   0.1041
gender1                 -1.150e-01  8.913e-01  1.702e-01 -0.676   0.4991
```

```
bas_age_gple30          -5.949e-02  9.422e-01  1.880e-01 -0.316   0.7517
bas_age_gpgteq40        -1.329e-01  8.756e-01  3.851e-01 -0.345   0.7301
sl_yn:bas_cd4_gpgteq200 -1.518e+01  2.555e-07  2.123e+03 -0.007   0.9943
sl_yn:bas_cd4_gp50_99   -1.749e+00  1.740e-01  8.796e-01 -1.988   0.0468 *
sl_yn:bas_cd4_gp100_199 -6.631e-01  5.152e-01  7.764e-01 -0.854   0.3931
sl_yn:bas_cd4_gple50    -9.423e-01  3.897e-01  5.096e-01 -1.849   0.0645 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                        exp(coef) exp(-coef) lower .95 upper .95
bas_cd4_gp50_99         4.466e+00  2.239e-01   2.47156    8.0709
bas_cd4_gp100_199       2.453e+00  4.077e-01   1.32276    4.5472
bas_cd4_gple50          7.219e+00  1.385e-01   3.90683   13.3403
adv_stage0              7.087e-01  1.411e+00   0.46794    1.0734
gender1                 8.913e-01  1.122e+00   0.63855    1.2442
bas_age_gple30          9.422e-01  1.061e+00   0.65179    1.3621
bas_age_gpgteq40        8.756e-01  1.142e+00   0.41160    1.8625
sl_yn:bas_cd4_gpgteq200 2.555e-07  3.913e+06   0.00000       Inf
sl_yn:bas_cd4_gp50_99   1.740e-01  5.748e+00   0.03103    0.9754
sl_yn:bas_cd4_gp100_199 5.152e-01  1.941e+00   0.11249    2.3599
sl_yn:bas_cd4_gple50    3.897e-01  2.566e+00   0.14354    1.0582

Rsquare= 0.002   (max possible= 0.047 )
Likelihood ratio test= 71.6  on 11 df,   p=6.057e-11
Wald test            = 58.61  on 11 df,   p=1.682e-08
Score (logrank) test = 74.21  on 11 df,   p=1.925e-11
```

Output for the interaction of advanced stage and second line (weighted)

```
                    coef exp(coef) se(coef)      z Pr(>|z|)
adv_stage0       -0.4305    0.6502   0.2235 -1.926  0.05404 .
gender1          -0.1150    0.8914   0.1702 -0.676  0.49919
bas_age_gple30   -0.0532    0.9482   0.1879 -0.283  0.77704
bas_age_gpgteq40 -0.1219    0.8852   0.3851 -0.317  0.75149
bas_cd4_gp50_99   1.5008    4.4851   0.3012  4.983 6.25e-07 ***
bas_cd4_gp100_199 0.9432    2.5681   0.3129  3.014  0.00257 **
bas_cd4_gple50    2.0216    7.5506   0.3097  6.528 6.65e-11 ***
sl_yn:adv_stage1 -1.5014    0.2228   0.5037 -2.981  0.00288 **
sl_yn:adv_stage0 -0.3184    0.7273   0.5922 -0.538  0.59083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
adv_stage0           0.6502     1.5380   0.41958     1.008
gender1              0.8914     1.1219   0.63853     1.244
bas_age_gple30       0.9482     1.0546   0.65610     1.370
bas_age_gpgteq40     0.8852     1.1297   0.41618     1.883
bas_cd4_gp50_99      4.4851     0.2230   2.48556     8.093
bas_cd4_gp100_199    2.5681     0.3894   1.39087     4.742
```

```
bas_cd4_gple50       7.5506     0.1324   4.11523    13.854
sl_yn:adv_stage1     0.2228     4.4880   0.08302     0.598
sl_yn:adv_stage0     0.7273     1.3749   0.22787     2.322

Rsquare= 0.002   (max possible= 0.047 )
Likelihood ratio test= 72.16  on 9 df,   p=5.739e-12
Wald test            = 63.11  on 9 df,   p=3.358e-10
Score (logrank) test = 73.06  on 9 df,   p=3.811e-12
```

B.1.4 Outputs for Chapter 5.2.5

Output for $S \sim T(t) + Wait(t) + \overline{B}$

```
                      coef exp(coef)  se(coef)      z Pr(>|z|)
sl_yn             -0.903379  0.405198  0.349773 -2.583 0.009801 **
total_wait         0.017364  1.017516  0.024747  0.702 0.482888
cd4               -0.004659  0.995351  0.001744 -2.672 0.007544 **
cd4lag1            0.001573  1.001574  0.001622  0.970 0.332286
gender1           -0.051196  0.950092  0.171929 -0.298 0.765875
bas_age_gple30     0.002404  1.002407  0.190924  0.013 0.989955
bas_age_gpgteq40  -0.046799  0.954279  0.387360 -0.121 0.903838
bas_cd4_gp50_99    0.921011  2.511829  0.345519  2.666 0.007685 **
bas_cd4_gp100_199  0.577398  1.781398  0.331368  1.742 0.081426 .
bas_cd4_gple50     1.336625  3.806176  0.367519  3.637 0.000276 ***
adv_stage0        -0.306419  0.736078  0.212195 -1.444 0.148728
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
sl_yn                0.4052     2.4679    0.2041    0.8043
total_wait           1.0175     0.9828    0.9693    1.0681
cd4                  0.9954     1.0047    0.9920    0.9988
cd4lag1              1.0016     0.9984    0.9984    1.0048
gender1              0.9501     1.0525    0.6783    1.3308
bas_age_gple30       1.0024     0.9976    0.6895    1.4573
bas_age_gpgteq40     0.9543     1.0479    0.4466    2.0389
bas_cd4_gp50_99      2.5118     0.3981    1.2761    4.9442
bas_cd4_gp100_199    1.7814     0.5614    0.9305    3.4105
bas_cd4_gple50       3.8062     0.2627    1.8521    7.8221
adv_stage0           0.7361     1.3586    0.4856    1.1157

Rsquare= 0.002   (max possible= 0.046 )
Likelihood ratio test= 78.29  on 11 df,   p=3.156e-12
Wald test            = 66.68  on 11 df,   p=5.204e-10
Score (logrank) test = 73.86  on 11 df,   p=2.244e-11
```

Output for $S \sim T(t) + Wait(t) + \overline{B}$ (weighted)

```
                       coef exp(coef) se(coef)      z Pr(>|z|)
sl_yn             -1.09454   0.33469  0.38108 -2.872  0.00408 **
total_wait         0.02129   1.02152  0.02558  0.832  0.40524
gender1           -0.12263   0.88459  0.17008 -0.721  0.47089
bas_age_gple30    -0.05292   0.94846  0.18793 -0.282  0.77826
bas_age_gpgteq40  -0.12267   0.88455  0.38503 -0.319  0.75002
bas_cd4_gp50_99    1.49337   4.45208  0.30116  4.959 7.10e-07 ***
bas_cd4_gp100_199  0.93435   2.54555  0.31286  2.986  0.00282 **
bas_cd4_gple50     2.01042   7.46643  0.31008  6.484 8.95e-11 ***
adv_stage0        -0.34665   0.70705  0.21185 -1.636  0.10177
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
sl_yn                0.3347     2.9878    0.1586    0.7064
total_wait           1.0215     0.9789    0.9716    1.0740
gender1              0.8846     1.1305    0.6338    1.2346
bas_age_gple30       0.9485     1.0543    0.6562    1.3708
bas_age_gpgteq40     0.8846     1.1305    0.4159    1.8813
bas_cd4_gp50_99      4.4521     0.2246    2.4673    8.0336
bas_cd4_gp100_199    2.5455     0.3928    1.3787    4.6999
bas_cd4_gple50       7.4664     0.1339    4.0661   13.7104
adv_stage0           0.7070     1.4143    0.4668    1.0710

Rsquare= 0.002   (max possible= 0.047 )
Likelihood ratio test= 70.6  on 9 df,   p=1.163e-11
Wald test            = 61.26  on 9 df,   p=7.666e-10
Score (logrank) test = 70.43  on 9 df,   p=1.253e-11
```

Output for $S \sim T(t) + Tot(t) + \overline{B}$

```
                       coef exp(coef)  se(coef)      z Pr(>|z|)
sl_yn             -1.192909  0.303337  0.558083 -2.138 0.032556 *
total_sl           0.029574  1.030016  0.047194  0.627 0.530887
cd4               -0.004690  0.995321  0.001756 -2.670 0.007582 **
cd4lag1            0.001557  1.001559  0.001637  0.951 0.341461
gender1           -0.045918  0.955121  0.172037 -0.267 0.789541
bas_age_gple30     0.003419  1.003425  0.190917  0.018 0.985713
bas_age_gpgteq40  -0.047706  0.953414  0.387345 -0.123 0.901979
bas_cd4_gp50_99    0.913317  2.492576  0.345777  2.641 0.008258 **
bas_cd4_gp100_199  0.573711  1.774841  0.331353  1.731 0.083377 .
bas_cd4_gple50     1.330789  3.784029  0.367634  3.620 0.000295 ***
adv_stage0        -0.308723  0.734384  0.212226 -1.455 0.145755
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
```

```
sl_yn                0.3033     3.2967    0.1016    0.9057
total_sl             1.0300     0.9709    0.9390    1.1298
cd4                  0.9953     1.0047    0.9919    0.9988
cd4lag1              1.0016     0.9984    0.9983    1.0048
gender1              0.9551     1.0470    0.6817    1.3381
bas_age_gple30       1.0034     0.9966    0.6902    1.4588
bas_age_gpgteq40     0.9534     1.0489    0.4462    2.0370
bas_cd4_gp50_99      2.4926     0.4012    1.2657    4.9088
bas_cd4_gp100_199    1.7748     0.5634    0.9271    3.3979
bas_cd4_gple50       3.7840     0.2643    1.8409    7.7783
adv_stage0           0.7344     1.3617    0.4845    1.1132

Rsquare= 0.002   (max possible= 0.046 )
Likelihood ratio test= 78.22  on 11 df,   p=3.26e-12
Wald test            = 66.73  on 11 df,   p=5.093e-10
Score (logrank) test = 73.78  on 11 df,   p=2.318e-11
```

Output for $S \sim T(t) + Tot(t) + \overline{B}$ (weighted)

```
                     coef exp(coef) se(coef)      z Pr(>|z|)
sl_yn             -1.34430   0.26072  0.61363 -2.191  0.02847 *
total_sl           0.02325   1.02352  0.05082  0.458  0.64727
gender1           -0.11894   0.88786  0.17015 -0.699  0.48453
bas_age_gple30    -0.05203   0.94930  0.18792 -0.277  0.78190
bas_age_gpgteq40  -0.12406   0.88333  0.38504 -0.322  0.74730
bas_cd4_gp50_99    1.49435   4.45644  0.30117  4.962 6.98e-07 ***
bas_cd4_gp100_199  0.93582   2.54930  0.31285  2.991  0.00278 **
bas_cd4_gple50     2.01437   7.49604  0.30993  6.499 8.06e-11 ***
adv_stage0        -0.34762   0.70637  0.21185 -1.641  0.10083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
sl_yn                0.2607     3.8355   0.07832     0.868
total_sl             1.0235     0.9770   0.92650     1.131
gender1              0.8879     1.1263   0.63608     1.239
bas_age_gple30       0.9493     1.0534   0.65682     1.372
bas_age_gpgteq40     0.8833     1.1321   0.41532     1.879
bas_cd4_gp50_99      4.4564     0.2244   2.46964     8.042
bas_cd4_gp100_199    2.5493     0.3923   1.38076     4.707
bas_cd4_gple50       7.4960     0.1334   4.08337    13.761
adv_stage0           0.7064     1.4157   0.46634     1.070

Rsquare= 0.002   (max possible= 0.047 )
Likelihood ratio test= 70.19  on 9 df,   p=1.396e-11
Wald test            = 61.17  on 9 df,   p=7.958e-10
Score (logrank) test = 70.28  on 9 df,   p=1.344e-11
```

B.1.5 Outputs for Chapter 5.3

Output for $S \sim T(t) + \overline{B}$ (normalized weights)

```
                       coef exp(coef) se(coef)      z Pr(>|z|)
sl_yn             -1.14136   0.31938  0.38280 -2.982  0.00287 **
gender1           -0.12387   0.88349  0.17034 -0.727  0.46709
bas_age_gple30    -0.05486   0.94662  0.18804 -0.292  0.77049
bas_age_gpgteq40  -0.12886   0.87909  0.38576 -0.334  0.73834
bas_cd4_gp50_99    1.49717   4.46902  0.30186  4.960 7.05e-07 ***
bas_cd4_gp100_199  0.93767   2.55402  0.31357  2.990  0.00279 **
bas_cd4_gple50     2.01480   7.49921  0.31074  6.484 8.94e-11 ***
adv_stage0        -0.34464   0.70847  0.21200 -1.626  0.10403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
sl_yn                0.3194     3.1310    0.1508    0.6763
gender1              0.8835     1.1319    0.6327    1.2337
bas_age_gple30       0.9466     1.0564    0.6548    1.3685
bas_age_gpgteq40     0.8791     1.1375    0.4127    1.8724
bas_cd4_gp50_99      4.4690     0.2238    2.4733    8.0752
bas_cd4_gp100_199    2.5540     0.3915    1.3814    4.7221
bas_cd4_gple50       7.4992     0.1333    4.0786   13.7885
adv_stage0           0.7085     1.4115    0.4676    1.0735

Rsquare= 0.002   (max possible= 0.046 )
Likelihood ratio test= 69.8  on 8 df,   p=5.387e-12
Wald test            = 60.89  on 8 df,   p=3.113e-10
Score (logrank) test = 69.91  on 8 df,   p=5.123e-12
```

B.2 Outputs for Chapter 6

B.2.1 Outputs for Chapter 6.1

Output for $S \sim T(t) + \overline{B}$ (unstabilized weights)

```
                        coef  exp(coef)   se(coef)      z Pr(>|z|)
sl_yn              -0.958536   0.383454   0.958873 -1.000   0.3175
gender1            -0.002347   0.997656   0.839004 -0.003   0.9978
bas_age_gple30      1.737256   5.681732   1.401710  1.239   0.2152
bas_age_gpgteq40    3.357025  28.703680   1.494596  2.246   0.0247 *
bas_cd4_gp50_99     4.268190  71.392311   2.731499  1.563   0.1182
bas_cd4_gp100_199   2.648081  14.126909   2.850797  0.929   0.3529
bas_cd4_gple50      5.400318 221.476845   2.766525  1.952   0.0509 .
adv_stage0          1.219427   3.385247   0.786270  1.551   0.1209
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                  exp(coef) exp(-coef) lower .95 upper .95
sl_yn                0.3835   2.607876   0.05855     2.511
gender1              0.9977   1.002350   0.19267     5.166
bas_age_gple30       5.6817   0.176003   0.36420    88.638
bas_age_gpgteq40    28.7037   0.034839   1.53368   537.205
bas_cd4_gp50_99     71.3923   0.014007   0.33775 15090.635
bas_cd4_gp100_199   14.1269   0.070787   0.05290  3772.683
bas_cd4_gple50     221.4768   0.004515   0.97827 50141.653
adv_stage0           3.3852   0.295399   0.72496    15.808

Rsquare= 0.001   (max possible= 0.002 )
Likelihood ratio test= 23.17  on 8 df,   p=0.003154
Wald test            = 19.39  on 8 df,   p=0.01289
Score (logrank) test = 38.74  on 8 df,   p=5.5e-06
```

B.3 Outputs for Appendix A

B.3.1 Outputs for Appendix A.2.1

Output for $Y(t) \sim T(t) + \overline{B}$

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1548  -0.0928  -0.0722  -0.0503   3.8993

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -6.38521    0.38984 -16.379  < 2e-16 ***
sl_yn             -0.92451    0.35035  -2.639  0.00832 **
gender1           -0.10709    0.17207  -0.622  0.53370
bas_age_gple30    -0.02393    0.19141  -0.125  0.90051
bas_age_gpgteq40  -0.10808    0.38814  -0.278  0.78065
bas_cd4_gp50_99    1.48459    0.30356   4.891 1.01e-06 ***
bas_cd4_gp100_199  0.93880    0.31454   2.985  0.00284 **
bas_cd4_gple50     1.96722    0.31381   6.269 3.64e-10 ***
adv_stage0        -0.32883    0.21257  -1.547  0.12188
ns(tstop, 5)1      0.06570    0.38661   0.170  0.86505
ns(tstop, 5)2     -0.52817    0.47283  -1.117  0.26398
ns(tstop, 5)3      0.01235    0.84670   0.015  0.98836
ns(tstop, 5)4     -2.70340    1.22739  -2.203  0.02763 *
ns(tstop, 5)5     -4.65756    2.27798  -2.045  0.04089 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1900.3  on 43865  degrees of freedom
```

```
Residual deviance: 1821.2  on 43852  degrees of freedom
AIC: 1849.2

Number of Fisher Scoring iterations: 10
```

Output for $Y(t) \sim C(t) + T(t) + \overline{B}$

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1766  -0.0925  -0.0685  -0.0486   4.2340

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -5.590344   0.448554 -12.463  < 2e-16 ***
cd4               -0.003278   0.000939  -3.491 0.000481 ***
sl_yn             -0.946496   0.350233  -2.702 0.006883 **
gender1           -0.046247   0.172394  -0.268 0.788498
bas_age_gple30     0.003016   0.191419   0.016 0.987430
bas_age_gpgteq40  -0.051517   0.388300  -0.133 0.894452
bas_cd4_gp50_99    0.879699   0.342884   2.566 0.010300 *
bas_cd4_gp100_199  0.549449   0.330023   1.665 0.095936 .
bas_cd4_gple50     1.293219   0.364535   3.548 0.000389 ***
adv_stage0        -0.305326   0.212639  -1.436 0.151032
ns(tstop, 5)1      0.295102   0.390013   0.757 0.449261
ns(tstop, 5)2     -0.195833   0.478895  -0.409 0.682593
ns(tstop, 5)3      0.615697   0.857859   0.718 0.472934
ns(tstop, 5)4     -1.954052   1.240324  -1.575 0.115156
ns(tstop, 5)5     -3.915071   2.288140  -1.711 0.087076 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1900.3  on 43865  degrees of freedom
Residual deviance: 1805.9  on 43851  degrees of freedom
AIC: 1835.9

Number of Fisher Scoring iterations: 10
```

Output for $Y(t) \sim T(t) + \overline{B}$ (with log(t) instead spline)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1749  -0.0916  -0.0725  -0.0514   3.8091

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -6.11426    0.34544 -17.700  < 2e-16 ***
```

```
sl_yn             -0.86838    0.34963  -2.484  0.01300 *
gender1           -0.11781    0.17204  -0.685  0.49349
bas_age_gple30    -0.01963    0.19130  -0.103  0.91827
bas_age_gpgteq40  -0.09094    0.38810  -0.234  0.81473
bas_cd4_gp50_99    1.44646    0.30341   4.767 1.87e-06 ***
bas_cd4_gp100_199  0.93115    0.31453   2.960  0.00307 **
bas_cd4_gple50     1.94199    0.31353   6.194 5.87e-10 ***
adv_stage0        -0.31967    0.21257  -1.504  0.13261
log(tstop)        -0.24623    0.07899  -3.117  0.00182 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1900.3  on 43865  degrees of freedom
Residual deviance: 1830.8  on 43856  degrees of freedom
AIC: 1850.8

Number of Fisher Scoring iterations: 9
```

B.3.2 Outputs for Appendix A.2.2

Output for $T(t) \sim \overline{B}$

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4316  -0.1487  -0.1085  -0.0770   3.8279

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.50650    0.20328 -22.169  < 2e-16 ***
gender1            0.04659    0.10659   0.437 0.662036
bas_age_gple30    -0.09131    0.11764  -0.776 0.437666
bas_age_gpgteq40  -0.33042    0.25673  -1.287 0.198072
bas_cd4_gp50_99    1.24211    0.18033   6.888 5.66e-12 ***
bas_cd4_gp100_199  0.89086    0.18151   4.908 9.20e-07 ***
bas_cd4_gple50     2.00064    0.18144  11.027  < 2e-16 ***
adv_stage0         0.13269    0.11697   1.134 0.256622
ns(tstop, 5)1     -1.16195    0.24495  -4.744 2.10e-06 ***
ns(tstop, 5)2     -1.69445    0.30854  -5.492 3.98e-08 ***
ns(tstop, 5)3     -2.33720    0.63066  -3.706 0.000211 ***
ns(tstop, 5)4     -3.75179    0.57494  -6.526 6.78e-11 ***
ns(tstop, 5)5     -2.31580    1.14375  -2.025 0.042894 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4245.4  on 39335  degrees of freedom
```

```
Residual deviance: 3901.4  on 39323  degrees of freedom
AIC: 3927.4

Number of Fisher Scoring iterations: 8
```

Output for $T(t) \sim \overline{B} + C(t)$

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4437  -0.1599  -0.1017  -0.0541   4.1278

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.4587583  0.3023332  -8.133 4.20e-16 ***
cd4               -0.0085706  0.0009844  -8.706  < 2e-16 ***
gender1            0.1253088  0.1066638   1.175 0.240074
bas_age_gple30    -0.0708607  0.1178454  -0.601 0.547639
bas_age_gpgteq40  -0.2543514  0.2570917  -0.989 0.322496
bas_cd4_gp50_99   -0.2208391  0.2330863  -0.947 0.343406
bas_cd4_gp100_199  0.0208298  0.1995900   0.104 0.916881
bas_cd4_gple50     0.3092404  0.2543068   1.216 0.223980
adv_stage0         0.1842096  0.1172778   1.571 0.116250
ns(tstop, 5)1     -0.7469944  0.2466994  -3.028 0.002462 **
ns(tstop, 5)2     -1.1060456  0.3107989  -3.559 0.000373 ***
ns(tstop, 5)3     -1.1536124  0.6374292  -1.810 0.070329 .
ns(tstop, 5)4     -2.2848957  0.5855766  -3.902 9.54e-05 ***
ns(tstop, 5)5     -0.8115597  1.1593211  -0.700 0.483908
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4245.4  on 39335  degrees of freedom
Residual deviance: 3785.9  on 39322  degrees of freedom
AIC: 3813.9

Number of Fisher Scoring iterations: 9
```

Output for $T'(t) \sim T(t) + \overline{B}$

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9596  -0.3560  -0.3145  -0.2803   2.7277

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.63625    0.08787 -30.001  < 2e-16 ***
sl_yn              0.40327    0.06140   6.568 5.11e-11 ***
```

```
gender1           -0.09144    0.04342  -2.106  0.03523 *
bas_age_gple30     0.06754    0.04833   1.398  0.16222
bas_age_gpgteq40   0.23149    0.09482   2.441  0.01463 *
bas_cd4_gp50_99   -0.41307    0.05857  -7.053 1.75e-12 ***
bas_cd4_gp100_199 -0.14605    0.05301  -2.755  0.00587 **
bas_cd4_gple50    -0.28764    0.07195  -3.998 6.40e-05 ***
adv_stage0         0.03728    0.04774   0.781  0.43480
ns(tstop, 5)1     -0.53616    0.10991  -4.878 1.07e-06 ***
ns(tstop, 5)2      0.47316    0.10850   4.361 1.29e-05 ***
ns(tstop, 5)3     -0.85307    0.14175  -6.018 1.76e-09 ***
ns(tstop, 5)4      0.28900    0.19344   1.494  0.13517
ns(tstop, 5)5      2.61690    0.16311  16.044  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18545  on 43865  degrees of freedom
Residual deviance: 18164  on 43852  degrees of freedom
AIC: 18192

Number of Fisher Scoring iterations: 6
```

Output for $T'(t) \sim T(t) + \overline{B} + C(t)$

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9410  -0.3555  -0.3141  -0.2776   2.7547

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.5028398  0.0948414 -26.390  < 2e-16 ***
sl_yn              0.3893093  0.0615133   6.329 2.47e-10 ***
cd4               -0.0005555  0.0001495  -3.716 0.000202 ***
gender1           -0.0738633  0.0436469  -1.692 0.090590 .
bas_age_gple30     0.0711819  0.0483411   1.472 0.140888
bas_age_gpgteq40   0.2287268  0.0948943   2.410 0.015938 *
bas_cd4_gp50_99   -0.5170229  0.0647113  -7.990 1.35e-15 ***
bas_cd4_gp100_199 -0.2197238  0.0564467  -3.893 9.92e-05 ***
bas_cd4_gple50    -0.3944981  0.0774068  -5.096 3.46e-07 ***
adv_stage0         0.0448605  0.0477905   0.939 0.347889
ns(tstop, 5)1     -0.4791262  0.1108946  -4.321 1.56e-05 ***
ns(tstop, 5)2      0.5506243  0.1103664   4.989 6.07e-07 ***
ns(tstop, 5)3     -0.7507165  0.1442814  -5.203 1.96e-07 ***
ns(tstop, 5)4      0.4370321  0.1972611   2.216 0.026726 *
ns(tstop, 5)5      2.7660066  0.1680276  16.462  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18545  on 43865  degrees of freedom
Residual deviance: 18149  on 43851  degrees of freedom
AIC: 18179

Number of Fisher Scoring iterations: 6
```

Output for $T''(t) \sim T(t) + \overline{B}$

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1321  -0.0890  -0.0719  -0.0557   3.9271

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.14552    0.48945 -14.599  < 2e-16 ***
sl_yn             -1.40973    0.46030  -3.063  0.00219 **
gender1           -0.15314    0.17987  -0.851  0.39454
bas_age_gple30     0.36440    0.21337   1.708  0.08767 .
bas_age_gpgteq40  -0.15451    0.48438  -0.319  0.74974
bas_cd4_gp50_99    0.71555    0.25419   2.815  0.00488 **
bas_cd4_gp100_199  0.45389    0.25606   1.773  0.07629 .
bas_cd4_gple50     0.83769    0.29562   2.834  0.00460 **
adv_stage0        -0.04787    0.20332  -0.235  0.81387
ns(tstop, 5)1      1.09923    0.48804   2.252  0.02430 *
ns(tstop, 5)2      1.29901    0.54882   2.367  0.01794 *
ns(tstop, 5)3      0.33610    0.73314   0.458  0.64664
ns(tstop, 5)4      0.07154    1.18473   0.060  0.95185
ns(tstop, 5)5     -1.30710    1.49774  -0.873  0.38282
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1761.5  on 43865  degrees of freedom
Residual deviance: 1717.6  on 43852  degrees of freedom
AIC: 1745.6

Number of Fisher Scoring iterations: 9
```

Output for $T''(t) \sim T(t) + \overline{B} + C(t)$

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1790  -0.0887  -0.0666  -0.0495   4.0329

Coefficients:
```

```
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -6.0511275  0.5335334 -11.342  < 2e-16 ***
sl_yn             -1.4541551  0.4598809  -3.162  0.00157 **
cd4               -0.0046291  0.0009255  -5.002 5.68e-07 ***
gender1           -0.0476764  0.1802908  -0.264  0.79144
bas_age_gple30     0.3990954  0.2136446   1.868  0.06176 .
bas_age_gpgteq40  -0.0927706  0.4843456  -0.192  0.84810
bas_cd4_gp50_99   -0.1169200  0.2930945  -0.399  0.68996
bas_cd4_gp100_199 -0.0891178  0.2721122  -0.328  0.74329
bas_cd4_gple50    -0.0971779  0.3443078  -0.282  0.77776
adv_stage0        -0.0044846  0.2035814  -0.022  0.98243
ns(tstop, 5)1      1.4096636  0.4901538   2.876  0.00403 **
ns(tstop, 5)2      1.7377278  0.5523062   3.146  0.00165 **
ns(tstop, 5)3      1.0841796  0.7424021   1.460  0.14419
ns(tstop, 5)4      1.0380419  1.1936385   0.870  0.38449
ns(tstop, 5)5     -0.3190660  1.5102110  -0.211  0.83267
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1761.5  on 43865  degrees of freedom
Residual deviance: 1685.2  on 43851  degrees of freedom
AIC: 1715.2

Number of Fisher Scoring iterations: 9
```

Output for $Y(t) \sim T(t) + \overline{B}$ (weighted)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5127  -0.0918  -0.0712  -0.0498   4.4557

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -6.38164    0.38883 -16.412  < 2e-16 ***
sl_yn             -1.21938    0.37877  -3.219  0.00128 **
gender1           -0.13430    0.17028  -0.789  0.43028
bas_age_gple30    -0.02835    0.18853  -0.150  0.88048
bas_age_gpgteq40  -0.11062    0.38682  -0.286  0.77490
bas_cd4_gp50_99    1.48487    0.30194   4.918 8.76e-07 ***
bas_cd4_gp100_199  0.94308    0.31319   3.011  0.00260 **
bas_cd4_gple50     2.01475    0.31084   6.482 9.07e-11 ***
adv_stage0        -0.33806    0.21208  -1.594  0.11094
ns(tstop, 5)1      0.08665    0.38418   0.226  0.82156
ns(tstop, 5)2     -0.47321    0.46741  -1.012  0.31135
ns(tstop, 5)3     -0.01327    0.81587  -0.016  0.98702
ns(tstop, 5)4     -2.36690    1.13412  -2.087  0.03689 *
ns(tstop, 5)5     -4.06057    2.06027  -1.971  0.04874 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1924.2  on 43865  degrees of freedom
Residual deviance: 1839.5  on 43852  degrees of freedom
AIC: 1862.2

Number of Fisher Scoring iterations: 10
```

Appendix C

Contents of the CD-Rom

C.1 Documentation

- `Pictures`: Folder of all the pictures used in the report.
- `Presentation`: Latex and PDF version of my power point presentation on the Thesis
- Others: All ".tex" files and the PDF version of the report

C.2 R Files

- `Extended Cox Model`: Folder of all ".R" files for the analysis with the extended Cox model
  - `Analysis_Cox_27.R`: Analysis of the sensitivity to model choice (Chapter 5.4) including the code for the plots of Chapter 5.4.
  - `Analysis_Cox_27_revised.R`: Analysis of the data set with the missing outliers (Chapter 5.5.1)
  - `Ananlysis_extended_Cox.R`: Analysis with the IPTW method and the Cox model
  - `Model_Selection_Cox.R`: Model selection for Chapter 5.4 with the Cox model
- `Pooled Logistic Regression`: Folder of all ".R" files for the analysis with the pooled logistic model (includes the same files as `Extended Cox Model` but with pooled logistic regression)
- `Prepare_Data.R`: Code to add the new variables to the existing data set

C.3 Papers

Includes the PDF versions of all the papers of the bibliography.