



**ATHENS UNIVERSITY OF ECONOMICS AND
BUSINESS**

DEPARTMENT OF STATISTICS

**Bayesian competing risks analysis of HIV data with
missing event types**

Ioannis Charalampopoulos

Thesis

submitted to the Department of Statistics of
the Athens University of Economics and Business
as part of the requirements for the acquisition
of the Master's Degree in Statistics

Athens

September 2023



Dedication

To my father and my mother because they are always encouraging me.



Acknowledgements

I sincerely thank Mr. Ioannis Ntzoufras, Professor at the Department of Statistics of Athens University of Economics and Business, Mr. George Bakoyannis, Assistant Professor of Statistics at the Department of Statistics in Athens University of Economics and Business and Mr. Constantin Yiannoutsos, Professor at School of Public Health of Indiana University, for their contribution throughout the duration of the postgraduate program, as well as for the constant help and encouragement they provided me, during the preparation of the Thesis as supervisors. Also, i would like to thank my family and loved ones for all the support they offered me throughout the preparation of my Thesis.



Abstract

HIV virus is a widely transmitted virus especially in Africa. Widely known as the longest pandemic nowadays, from 1981, it is responsible for the death of over 40 million people. It is crucial, for the health community, to identify the most influential factors of transmission as well as the mortality rate and the groups of individuals that are more vulnerable to lose their lives. Also, it is very important to identify, in the presence of multiple possible outcomes of interest, the outcome that will be expressed in the patients in order suitable actions to be implemented. Precise results will assist the health community to act highly effective and save as more lives as possible.

In this thesis, the definition of survival analysis and competing risks will be introduced, along with methods to estimate possible missing outcomes and a brief analysis of the dataset will be given. It is very important to be mentioned that there are two possible outcomes of interest in this particular dataset, loss of life and disengagement from care. Furthermore, different models will be fitted in both frequentist and Bayesian frameworks and different approaches will be given for the estimation of the outcome of interest for patients that have been wrongly classified as disengagements.



Contents

1	Survival Analysis in medical research	1
1.1	A definition of survival analysis	1
1.2	Some characteristics of survival analysis and survival data	2
1.3	Important functions in survival analysis	3
1.4	Non-Parametric approaches	4
1.5	Tests for the comparison of survival groups	7
1.6	Explanatory variables in the survival analysis	8
1.7	Weibull and exponential distributions in the survival analysis	10
1.8	Introduction to competing risks	12
1.9	Cause-specific hazard function and Cumulative incidence function	13
1.10	Aalen-Johansen estimator	16
1.11	Inference in the presence of competing risks	17
1.12	An example for the Fine and Gray model	18
1.13	Conclusion of the first chapter	21
2	Dealing with misclassification and missing event types in competing risk settings	23
2.1	General idea and description of the problem and a Bayesian definition	23
2.2	The two possible approaches to tackle the issue	27
2.3	Approaches for the estimation of the mortality of the disease and the probability of a cause of failure to take place	30
2.4	Augmented Inverse Propensity Weighted estimators	31
2.5	Multiple Imputation approach	32
2.6	Important notation for the approach in scope	33
2.7	A toy example	35
2.8	Important functions for the implementation of MPPL	37
2.9	The constructed models of the frequentist's statistics	39



2.10	Conclusion of the second chapter	40
3	Description of the disease, the dataset and creation of the models in scope	41
3.1	Description of the disease	41
3.2	Description of the dataset	42
3.3	Pairwise comparisons between the variables	52
3.4	Complete case analysis with Cox proportional hazards model	55
3.5	Multiple Imputation and Cox model using the predicted probabilities of logistic regression	60
3.6	MPPLE approach	65
3.7	Comparison of the models and conclusion of third chapter	70
4	Bayesian approach for competing risk data with missing event types	75
4.1	Selection of the prior distribution	75
4.2	Complete case analysis using the Weibull baseline hazard function	77
4.3	Imputation of the missing outcomes using Bayesian logistic regression and the Weibull baseline hazard function	83
4.4	Comparison of the models and conclusion of fourth chapter	87
5	Conclusion and further discussion	91
A	Appendix	95





Chapter 1

Survival Analysis in medical research

1.1 A definition of survival analysis

Survival analysis plays a significant role in the medical research. In general, survival analysis is a collection of statistics that investigates the time to an event, i.e., loss of life for a group of people, disease recurrence, or failure of a system. Using specific procedures, scientists are trying to predict the mortality of a rare disease, better understand the outcome of a treatment or assess the effectiveness of treatments. Scientists employ various statistical techniques and models within survival analysis to analyze and interpret the survival data. These techniques include Kaplan-Meier estimation, Cox proportional hazards model, and parametric survival models, among others. Furthermore, it is certain that death as the end-point of the analysis is the central point for the majority of the research. However, loss of life is not the only event where survival analysis can be used. Examples like the lifetime of an electronic device, the vaccination effectiveness or the time taken by an individual to complete a task in a psychological study are very common. In this thesis, loss of life will be one of the possible outcomes that a patient can develop. This aforementioned outcome will be considered as a competing risk, namely it will “compete” other possible outcomes. More regarding competing risks, on the second chapter of the thesis.



1.2 Some characteristics of survival analysis and survival data

One common characteristic of survival data is the positive skewness. There is no symmetrical distribution of the data and since we are studying time to death data, it is reasonable that some individuals will “last longer” than some others. Normal distribution is not the best option to use because of the longer right tail that the data have. Popular distributions in the survival analysis are the exponential and the Weibull distributions.

Another common and very significant feature is censoring. We say that a patient is censored when the event of interest has not been observed, i.e., when an individual enters a medical study and after a period of time in the study, she or he has been lost from the program (if competing risks are not in place) or the study has come to an end. The last example is called administrative or fixed censoring. There are three types of censoring: Right, left and interval censoring. Right censoring is the most common and occurs when the individual does not express the event of interest in the observed time of the study and their last follow-up time is less than their time to their possible death. In this thesis, we will consider that right censoring is applied.

A very significant assumption that censored survival data should follow has to do with independent censoring. Assuming that a group of individuals has similar characteristics, namely have similar health history or the same age, then the subjects whose times are censored must be representative of all the subjects that are still at risk. In other words, even if we don't know the actual time to event for a patient, we could assume that it should follow the logic of those patients who are still in the study. However, it's worth noting that the assumption of independent censoring should be carefully evaluated in practice, as violations of this assumption can affect the validity of the results.



1.3 Important functions in survival analysis

There are three functions that play significant role in the analysis of the survival data. These are the **survivor function**, the **hazard function** and the **cumulative hazard function**. In these functions, we will use the notation T , which is a random variable and counts the time for the individual to express the outcome of interest.

a) Survivor function

$$S(t) = P(T \geq t) = 1 - F(t) \text{ for } t > 0, \quad (1.1)$$

where $F(t)$ is the distribution function of T . The survivor function represents the probability that an individual survives beyond time t or the event of interest that we are looking for has not yet occurred up to that time.

b) Hazard function

$$\lim_{h \rightarrow 0} \frac{P(t \leq T < t+h | T \geq t)}{h} = \frac{f(t)}{S(t)}. \quad (1.2)$$

Hazard function is used in order to express the “intensity of death” in the interval $(t, t+h)$ for an individual who has survived till time t . Function $h(t)$ is also referred as the instantaneous death rate or the force of mortality. It is an important concept in survival analysis for the understanding of the underlying risk factors and changes in risk over time.

c) Cumulative hazard function

CHF measures the total amount of risk of death happening by time t . The CHF can only increase or remain the same. The cumulative hazard function, which is



frequently employed in parametric survival models and other survival analysis techniques, offers useful insights into the cumulative risk profile of the event of interest. From the equation 1.2, it follows that

$$h(t) = -\frac{d}{dt} \log S(t) \quad (1.3)$$

and if we do some calculations, we take

$$S(t) = e^{-H(t)}, \quad (1.4)$$

where

$$H(t) = \int_0^t h(u) du \quad (1.5)$$

1.4 Non-Parametric approaches

It should be mentioned that a parametric approach is not a lot of times the best solution, namely when the distribution of survival times is not well-known or the assumptions about the distributional form are not reasonable. Non-parametric techniques for modelling survival data can be used like the well-known Kaplan-Meier estimate and the Life-table estimate. These estimates do not require the assumption that data follow a specific distribution.

The life-table method is a method widely used by actuaries to solve problems like life insurance rates or requisite reserves. It is a widely used method in survival analysis (not like Kaplan-Meier) since it provides a serviceable method to analyze survival data. Assume that we have observations divided into a series of time intervals. Our goal is to estimate the survival function. Let d_j be the number of deaths, c_j the number of censorings in the interval and n_j the number of individuals that are still alive. The life-table estimate of the survivor function is the following:

$$\hat{S}(t) = \prod_{i=1}^{j-1} \left(\frac{n_i^* - d_i}{n_i^*} \right), \quad (1.6)$$



, where $n_j^* = n_j - \frac{c_j}{2}$.

The main assumption is that in the span of the interval, deaths or censorings are happening randomly in the pre-ordered intervals that we have set. The intervals are usually split equally, but this is not necessarily the case. Life-table method calculates the survival at each interval and because of the fact that the censorings are random, it is possible that certain intervals will have a lot of deaths and censorings and other intervals will have no deaths or censorings. That was the reason, a better approach to be utilized such as Kaplan-Meier.

Kaplan-Meier is considered as a variant of the life table method. In this method the intervals are not fixed by length but they are defined by the occurrence of an event. In other words, one death time is observed in each of the intervals and this observation will occur at the start of the interval. The Kaplan-Meier method follows three assumptions. Firstly, as it has already been discussed, the occurrence of an event happens at a specified time. Secondly, if a censoring and a death occur at the same time, then we consider that death happens first and the censoring immediately after. Last but not least, we assume that the reason an observation is censored does not relate to the cause of failure.

Regarding the survival probability of the Kaplan-Meier method, if we assume, like in the life table method, that n_j is the number of individuals in the study at time t_j , d_j is the number of deaths and c_j the number of censorings at that time, then the cumulative survival probability for all the time intervals is:

$$\hat{S}(t) = \prod_{j=1}^k \hat{p}_j = \prod_{j=1}^k \left(\frac{n_j - d_j}{n_j} \right), \quad (1.7)$$

The survival probability $\hat{S}(t)$ has the value 1 at the start of the study and only changes after a patient loses her/his life. There is no change when a censoring is happening and the value goes to 0 if the last observation is uncensored. This is portrayed in the arbitrary toy example of the Kaplan-Meier curve below (please refer to Figure 1.1), where we simulated the time of the event and the outcome (death or censoring) of 15 patients. We assumed that the time follows a discrete uniform distribution, whereas the events of death and censorings follow an exponential distribution with the difference that death events are simulated with a slightly bigger rate. It is obvious that the estimated survival probability is 1 at the start of the observation period and reduces progressively when a death occurs. As mentioned above, censorings do not change the survival probability and, in figure 1.1, are portrayed with small crosses. Finally, because the last dead patient was in the week 74 and there were two more whose outcome was censored, the survival

probability did not go to zero.

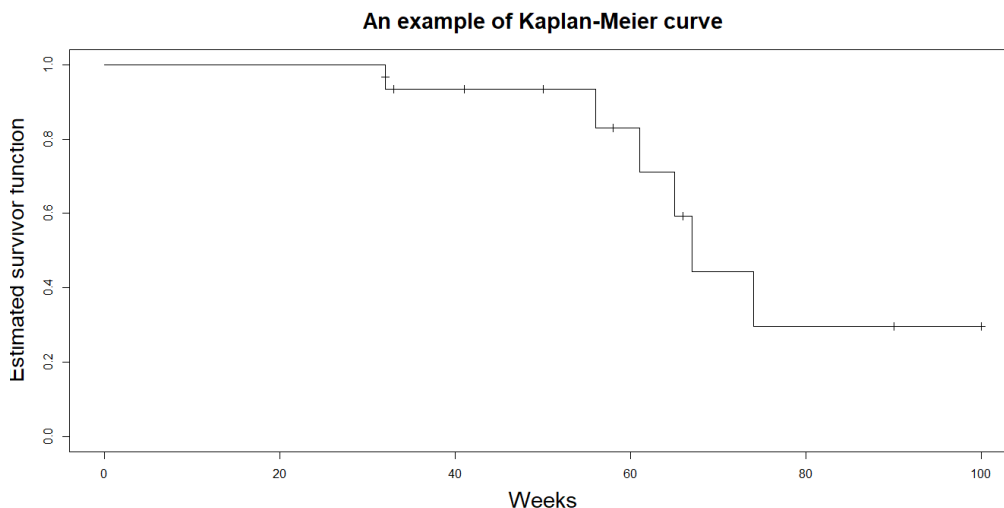


Figure 1.1: An arbitrary example of Kaplan-Meier curve.

Except of the Kaplan-Meier estimator, Nelson-Aalen estimator is also commonly used. This estimator's point of reference is the subject's event time and the formula for the estimation of the cumulative hazard function is

$$\hat{H}(t) = \sum \frac{d_j}{n_j}, \quad (1.8)$$

where we assume that d_j and n_j are the same as in the Kaplan-Meier method. As it is already known, the connecting relationship between survivor and hazard function is $S(t) = e^{-H(t)}$, so the estimation of the survivor function using the Nelson Aalen method is

$$\hat{S}(t) = \prod_{j=1}^k e^{-\frac{d_j}{n_j}}, \quad (1.9)$$

If the two methods will be compared, it is clear that at any given time for $x \geq 0$, the Nelson-Aalen estimate of the survivor function will exceed the respective estimate of the Kaplan Meier since in general, $e^{-x} \geq 1 - x$. Furthermore, despite the fact that Nelson-Aalen usually performs better than Kaplan-Meier when small

samples is the case, it is preferable to use the Kaplan-Meier since it is considered to be closer to the empirical survivor function (*Collet, 2003*). In general, both estimators have their merits and are commonly used in different scenarios, depending on the specific characteristics of the data and research objectives.

1.5 Tests for the comparison of survival groups

One of the most significant parts of survival analysis is without doubt the comparison between two groups. If we take for example two groups where the first group receives the standard treatment and the second the new treatment, we want to examine if the new treatment is better or worse than the one that we already have. It is very important for an analysis to use hypothesis tests in order to test whether there are differences, regardless the size, between the existed groups. The two well-known non-parametric procedures are the log-rank test and the Wilcoxon test.

Before defining what are these tests, we should set the null and the alternative hypotheses regarding the two groups of survival data. Following our example, the null hypothesis assumes that there is no statistical difference or existed relationship between the two groups that we compare, namely the new treatment is not better or worse than the custom. On the other hand, the alternative hypothesis assumes that a difference exists and must be taken into consideration.

The log-rank test examines whether there are differences in the survival curves between two (or more independent groups). This test is better to be used when the survival curves do not cross each other, because if they are used when they are overlapping, the test loses power. In general, the log-rank test is more powerful under the assumption of proportional hazards between the groups. The goal is to assess if the null hypothesis is valid. The creation of a test statistic will help with the solution of this problem. The log-rank statistic is obtained by the formula

$$T(t) = \frac{U}{V} = \frac{\sum_j (d_{1j} - E(d_{1j}))}{\sum_j \text{Var}(d_{1j})}, 1 \leq j \leq r, \quad (1.10)$$

where d_{1j} are the deaths at time $t(j)$ for the first group. U is the sum of the differences between the actual deaths in one of the groups and the expected deaths



under the assumption of the null hypothesis. It does not matter which of the two groups is used in the calculations because the result will be the same. V is the variance of U , which is the sum of the variances of d_{1j} because of the fact that a death time is independent with other death time. If the value of the test statistic T is large, then there is evidence that the null hypothesis could be rejected.

The Wilcoxon test is also used for the same reason as the log-rank test. However, the difference is that the log-rank test is considered to be more appropriate to identify distributional differences later in the follow-up time, where less individuals are investigated. On the other hand, Wilcoxon test performs better in identifying differences in the beginning of the analysis (*Lee, Desu, Gehan, 1975; Prentice and Marek, 1979*). Furthermore, as it has been referred above, when there is overlapping between the survival curves of two groups, the Wilcoxon method is considered to be a better solution. Regarding the test statistic, the difference here is in the formula:

$$U = \sum_j n_j (d_{1j} - E(d_{1j})), 1 \leq j \leq r, \quad (1.11)$$

where the difference between the deaths is weighted by n_j , namely the total number of individuals at risk at time $t(j)$.

1.6 Explanatory variables in the survival analysis

In survival analysis, explanatory variables like age, sex or specific habits (e.g., smoking, sexual life) are very important components regarding our understanding of a fatal disease or a new treatment. However, methods like Kaplan-Meier and life-table are not the most suitable when the case is to assess the influence of specific categorical variables. We need to find a modelling approach where we can assess the possible effect of the explanatory variables in the hazard function. Furthermore, as it was mentioned above, the log-rank test is suitable when the assumption of the proportional hazards is valid, namely if we have two groups then the hazard for an individual in the one group is proportional to the hazard of an individual in the other group. In other words, the hazard ratio for the two individuals is constant over time (let ψ).



Continuing our example, let $h_0(t)$ be the hazard function for an individual in the normal treatment and $h_2(t)$ the hazard function for an individual in the new treatment. If the aforementioned assumption regarding the proportional hazards is valid then we have $h_2(t) = \psi h_0(t)$. Since the hazard ratio is always positive, the formula $h_2(t) = \psi h_0(t)$ can be written as $h_2(t) = e^{\beta X} h_0(t)$ (a). The hazard ratio can give a very good understanding regarding the possible outcome of the new treatment in comparison to the standard one. The last formula is also known as the Cox regression model, where β is the vector of the regression coefficients of the explanatory variables and X is the vector of the values of the explanatory variables. It is a model that combines the non-parametric baseline hazard function with the covariate effects of $e^{\beta X}$.

The $h_0(t)$ can be denoted as the baseline hazard function, which is analogous to the intercept term in linear regression. In the baseline hazard function $h_0(t)$ all the covariates of the model are zero. Moreover, the exponent in the formula (a) is the linear predictor, namely for k coefficients we get

$$n_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ik}. \quad (1.12)$$

So, with the expansion of the equation (a), we will obtain the final equation

$$h_2(t) = h_0(t) e^{(\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ik})}. \quad (1.13)$$

The Cox regression model is considered as semi-parametric, namely a model with parametric and non-parametric components. Even if we know the regression parameters, the baseline hazard function $h_0(t)$ is unspecified and as a result the distribution of the model is not known.

Regarding the betas in the Cox regression model, these can be estimated thanks to the method of maximizing the partial likelihood. The partial likelihood for Cox models is:

$$L(\beta) = \prod_{i \in D} \frac{e^{Z_i^T \beta}}{\sum_{k \in R(t_i)} e^{Z_k^T \beta}}, \quad (1.14)$$

where D is the index set of death times and R the index set of subjects at risk at time t . In general, this is the observed likelihood for the subject i . It is clear that this formula is independent of the h_0 . Last but not least, the partial likelihood depends on the order of the events. Partial likelihood is used instead of the fully likelihood function, due to semi-parametric nature of the model. Moreover, the partial likelihood is used due to the fact that simplifies the estimation process



and the results, when complex or unknown baseline hazards are in place, can be considered as robust.

These estimations are very useful since they can help to find an estimate of the baseline hazard function and of course to point out a possible explanatory variable in the study (i.e., age or habits). Supposing that betas have been estimated using maximum likelihood or other methods, confidence intervals for these parameters will play significant role giving insight regarding the hypothesis. Specifically, the $100(1 - \alpha)\%$ confidence interval with limits $\hat{\beta} \pm \frac{Z_{\alpha}}{2} * s.e((\beta))$ will give insight with the existence or not of zero inside the interval. Assuming that zero is not inside, this is evidence that an explanatory variable is statistically significant and there may be a difference between the treatments in scope.

The Wald test is a widely used process that draw conclusions regarding the significance of the parameter β and is used for asymptotically normally distributed estimators. The statistic is the fraction

$$\frac{\hat{\beta}}{s.e(\hat{\beta})}, \quad (1.15)$$

where $s.e(\hat{\beta})$ is the standard error of $\hat{\beta}$. Furthermore, under the assumption of the null hypothesis, Wald test follows an asymptotic standard normal distribution. If the obtained p-value is not smaller than the defined limit, then the assumption of null hypothesis, namely that there is no difference with the presence of a specific explanatory variable, seems not to be followed.

1.7 Weibull and exponential distributions in the survival analysis

As we referred above, the two most important parametric distributions are the exponential and the Weibull. The Weibull distribution is a generalization of the exponential. The difference is that the exponential distribution assumes that the hazard function does not change over time ($h(t) = \lambda$), while the Weibull distribution is more flexible since the hazard function is not constant over time ($h(t) = \lambda \gamma t^{\gamma-1}$ where γ the shape parameter and λ the scale parameter).



The constant hazard function of the exponential distribution is the result of the memory-less property. This property basically tells us that the foreknowledge that an event has not occurred until time t does not influence the probability of the event to occur or not in the future. This is the reason why the use of the exponential is not the best, since the failure probability remains constant with the passing of the years.

The probability density function of the exponential distribution is $f(t) = \lambda e^{-\lambda t}$ while the survivor function is $S(t) = e^{-\lambda t}$. Assuming n observations, the likelihood for the exponential distribution is

$$L(\lambda) = \prod_{i=1}^n f(t_i)^{\Delta_i} S(t_i)^{1-\Delta_i} = \lambda e^{-\lambda t_i^{\Delta_i}} (e^{-\lambda t_i})^{1-\Delta_i}, \quad (1.16)$$

where Δ_i is one if the event of the i individual has occurred and zero if the survival time is censored. The goal is the estimation of λ . After some calculations in the formula 1.16, we get the expected value

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^n t_i}, \quad (1.17)$$

for which the log-likelihood maximizes itself. Moreover, r is the number of total deaths in the analysis. The mean of the exponential distribution is $\mu = \lambda^{-1}$, so the expected value for the mean is

$$\hat{\mu} = \frac{\sum_{i=1}^n t_i}{r}, \quad (1.18)$$

In other words, the mean survival time is the ratio of the total time the n persons spent alive to the number of deaths recorded. Based on this value, confidence intervals can be constructed for a better understanding. On the other hand, the probability density function of the Weibull distribution is

$$f(t) = \lambda \gamma t^{\gamma-1} e^{-\lambda t^\gamma}, \quad (1.19)$$

and the survivor function is

$$S(t) = e^{-\lambda t^\gamma}. \quad (1.20)$$

Following the same procedure as previously, the likelihood for n observations is

$$L(\lambda) = \prod_{i=1}^n (\lambda \gamma t_i^{\gamma-1} e^{-\lambda t_i^\gamma})^{\Delta_i} (e^{-\lambda t_i^\gamma})^{1-\Delta_i}. \quad (1.21)$$



After some calculations, the obtained estimation for λ is

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^n t_i \hat{\gamma}}, \quad (1.22)$$

while the estimation for the value γ can be estimated using numerically iterative procedures like Newton-Raphson.

1.8 Introduction to competing risks

So far, we have considered that there is only one type of failure, i.e., death. The idea of competing risks is that there is more than one possible outcomes during the follow-up study. Imagine that we study breast cancer in a group of people and we want to determine whether they will develop the event of interest. It is natural that the subject can also die from other causes like an accident or stroke. These events are referred as competing risks, since they are “competing” each other in order to become the event of interest. However, the result of a competing event is not necessarily death. For example, a study takes place in order to examine the prevalence of a disease and one of the subjects receives the vaccine against that disease. As a result, the vaccination behaves as a competing event.

It should be noted that if a subject dies from a competing risk like a stroke, she/he cannot die again from the primary event of interest like breast cancer, since this event is eliminated. These events are considered to be mutually exclusive and as a result they cannot happen simultaneously. Though, there are cases where the mutuality is not necessary. For example, if in a group of people that they receive a specific medicine or therapy, doctors decide to change the therapy, these act as competing risks but these events are not mutually exclusive. Furthermore, competing risks should not be treated as censoring events, since the assumption of random censoring could be violated.

Furthermore, censoring has as a result to overestimate the cumulative incidence estimators. For example, a subject that expresses an outcome other than the primary (i.e., a stroke in our example) will have a different prognosis than a subject who has not developed an event at all. A possible explanation can be that the former may be older or has more stress than the latter. As a result, these cases



should not coincide. Moreover, censoring the competing events will create biased cumulative incidence estimators.

One very significant question that needs to be answered is if the competing risks are independent from each other. If information regarding a subject's risk of experiencing one type of event does not reveal information about the subject's risk of experiencing the other type of event, the two competing risks are said to be independent. However, there is not yet a formally test in order to examine if two competing events are considered to be independent of one another.

There are applications in medical studies, where dependence is considered between the competing events. For example, a subject that has the possibility to die from a cardiovascular disease, has also the possibility to die from a different cause of death if he/she is in an advanced age. In many survival studies age is considered as a very common risk factor. Of course, age is not the only common risk factor. Physical inactivity or hereditary health problems can be significant risk factors. However, there are factors that can be protective to the subject regarding the competing risk, like healthy nutrition, a new drug or even the location.

1.9 Cause-specific hazard function and Cumulative incidence function

The cause-specific hazard function is very useful in the context of competing risks. It generalizes the traditional idea of the hazard function to more than one outcome of interest. Let C be the type of the outcome that has happened ($C=1, \dots, n$) and let T be the time the first event has occurred. Then given that the subject has not yet encountered any outcome of type i , the instantaneous rate at which an event of type I will occur is:

$$h_j^{cs}(t) = \lim_{h \rightarrow 0} \frac{P(t \leq T \leq t+h, C = j | T \geq t)}{h}, j = 1, \dots, n \quad (1.23)$$

As a result, the overall survival function, namely the probability to remain event-free until time t , is $S(t) = e^{-\sum_{j=1}^I H_n(t)}$, where $H_n(t)$ is the cumulative cause-specific hazard function for the n th event type:



$$H_n(t) = \int_0^t h_t(s) ds. \quad (1.24)$$

One major difference between the simple survival analysis, that has one event, and survival analysis with competing risks, is the fact that the one-to-one connection between the cumulative incident function and the cause specific hazard is no longer applicable. The Cumulative incidence function is the probability that an individual will survive until time t from a specific cause, let j . Then,

$$F_j(t) = P(T \leq t, C = j) = \int_0^t h_j(u) S(u) du. \quad (1.25)$$

Depending on the reason of the work, it depends whether we will use the cause specific hazard function or the cumulative incidence function. If the goal is to make prediction, the latter is more appropriate. CIF is extremely useful in clinical decision making and in the evaluation of a population's response to an intervention in terms of its effectiveness from a public health perspective. If the goal is to study the causality of a disease (i.e., evaluation of risk factors for various outcomes), then the former is better. However, it should be noted that for a better understanding of the problem in scope, a combination of the two functions is preferable.

Next, a new arbitrary toy example will be implemented in order to visualize the estimated probability of an event to happen. Suppose that there are 50 patients that need to be examined whether they will die from a cardiovascular disease. These patients can either die from a cardiovascular disease or a non-cardiovascular disease or even from an accident. Using the cuminc function, we plot and create the cumulative incidence analysis where we can see the probabilities of each of the possible events in the presence of the others through time.



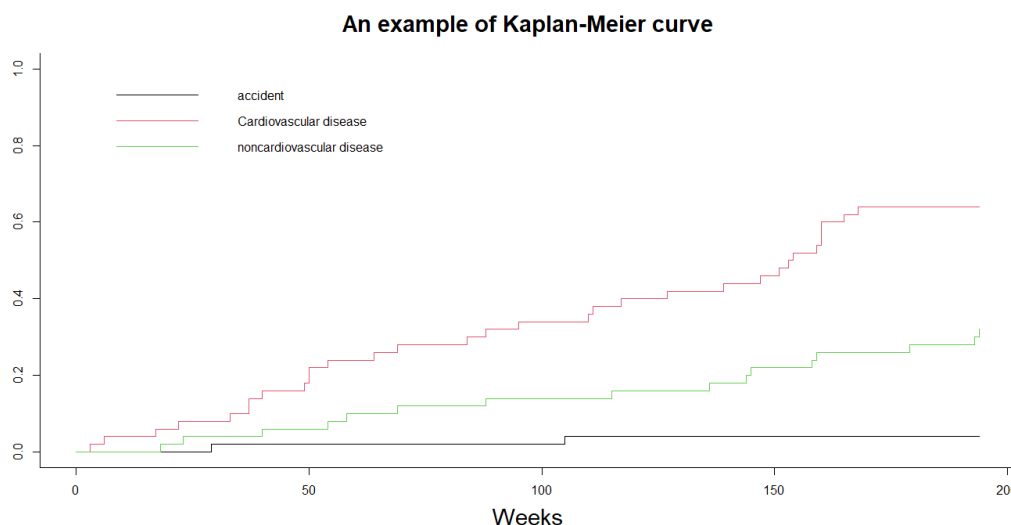


Figure 1.2: An arbitrary example of cumulative incidence analysis in the presence of competing risks. The curves of the three competing events are depicted in the plot.

It should be noted that the example was created with simulation. The time follows a discrete uniform distribution and the three competing events follow the exponential distribution with different rates, depending on the probability these events to take place. As we can see in Figure 1.2, the probabilities increase through the follow-up time and because of the rates that were used in the example, that was implemented in the R software, the probability a subject to die from a cardiovascular disease is increased. Moreover, these probabilities should sum to one if time is supposed to reach infinity.

Kaplan-Meier would be a way to estimate the survivor function. However, there are drawbacks regarding this estimate when competing risks is the case. First of all, if competing risks are in place, then the estimates from Kaplan-Meier are naive, since there are circumstances where the estimates are probabilistic impossible (sum of the estimated probabilities larger than one). Furthermore, from the definition of Kaplan-Meier, it is known that the estimators do not take into account the chance that a certain outcome may never take place because they are based on the assumption that an individual will eventually pass away from any given cause. Moreover, the estimator tends to overestimate the likelihood of a

specific occurrence.

As we referred above, the one-to-one correspondence between the CSH and CIF is no longer applicable. This fact creates some serious problems for the estimates. Depending on the effect, there are different cause-specific hazards. First of all, in the presence of competing risks, the cumulative incidence and hazard functions cannot be estimated from one and only model. This is because one covariate may influence in a different way the two functions of interest for a cause j . Furthermore, if the cumulative incidence function is the case, then the estimator is assumed naive, since deaths are considered the topic of interest. Every other cause of failure is considered as a censored observation.

From the formula of the survivor function, we get the naive estimator of the survivor function:

$$1 - \hat{S}(t) \quad (1.26)$$

With the assumption regarding the censored observations, 1.26 can be considered as one minus the Kaplan-Meier estimator, that is used to estimate risk functions. This estimator is considered as biased, since 1-Kaplan-Meier overestimates the cumulative incidence if competing risks are in place (*Sarah Lancy et al. (2016)*). As a result, the interpretation is not interpretable.

1.10 Aalen-Johansen estimator

A very good alternative is the Aalen-Johansen estimator, since it works better than 1-Kaplan-Meier in the presence of competing risks. It is a non-parametric estimation of the risks and it is unbiased. Firstly, as it is already known, $S(t)$ is the Kaplan-Meier estimator of the overall survivor function. Let d_{ij} be an indicator that takes the value one if the i event takes place for the j subject and zero for all the other cases. Moreover, let $R(t_j)$ be the individuals that are still alive at the moment before time j . Then the Aalen Johansen estimator for cause 1 is

$$\hat{F}_1(t) = \sum_{t_j \leq t} \hat{S}(t_j - 1) \frac{d_{1j}}{R(t_j)} \quad (1.27)$$



where the fraction is the Nelson-Aalen estimator that was mentioned in the formula 1.8. Comparing the Aalen-Johansen estimator with the 1-Kaplan-Meier, it is known that

$$F_1(t) \leq 1 - S_1(t) \quad (1.28)$$

Using $1 - S_1(t)$, there is overestimation of the cause 1 risk. This is not happening if the Aalen-Johansen estimator is utilized. The degree of bias is proportional to the magnitude of the competing risk. The higher the frequency of the competing risk, the more inappropriate the choice of 1-Kaplan Meier.

1.11 Inference in the presence of competing risks

In the analysis of competing risks, inference for these data is very important. Regarding the cumulative incidence function, if the goal is to make formal comparisons between two groups, one of the best solutions is the Gray's test (*Gray, 1988*). The Gray's test is a two-sample test and it is employed to determine whether the cause-specific cumulative incidence functions between two groups are equal and is considered as a modification of the log-rank test for competing risk data. The test is actually used in order to compare sub-distribution hazards (cause-specific cumulative incidence functions).

In the survival analysis with one event of interest, simple Cox models played a very significant role in the inference. However, in the case of competing risks the situation is a little bit different. The “match-up” between the rate and the risk is no longer possible and as a result a covariate influences the rate in a different way from how it affects the risk. As it has already been mentioned, in the “simple” survival with one possible outcome, the hazard function can easily give the survivor function. In the presence of the competing risks other solutions should be followed in order regression models for the cumulative incidences to be created. The most known model in order to acquire knowledge for the cumulative incidence is the Fine and Gray competing risks model, which is considered as a modification of the Cox model.

The Fine and Gray competing risks semi-parametric proportional hazards model provide a more accurate assessment of the effect of risk factors for a particular



event when competing hazards is the case. It is a mathematical way that can be used in order to create the one-to-one relationship that is no longer applicable due to competing risks. With the use of that model, it is easier to model the cumulative incidence. In general, a pitfall regarding the Fine and Gray test is that in some situations and for specific time-slots, the sum of the probabilities for various event types may exceed one (*Peter C. Austin, 2022*).

Another drawback of this method is that the parameter estimates are not easily interpretable. The hazard function for the j th cause is

$$\lambda_j^*(t) = \frac{1}{1 - F_j(t)} \frac{dF_j(t)}{dt} \quad (1.29)$$

where $F_j(t)$ is the cause-specific cumulative incidence function for the cause j . In that specific method, subjects that are not actually at risk from a cause j , because they have failed from other reason, are also included in the group of individuals that are at risk. The overall hazard $\lambda(t)$ is the sum of the cause-specific hazards for every event of interest. Generalizing formula (a) in the competing risks we get

$$\lambda_{ij}(t) = \lambda_{0j}(t)e^{\beta_j X_i} \quad (1.30)$$

where λ_{ij} is the instantaneous hazard of the cause j for the subject i . Furthermore, $\lambda_{0j}(t)$ is the corresponding baseline hazard function for the cause j , while x_i is the vector of the explanatory variables for the subject i . Naturally, if there are k different competing events, k possible different models can be created.

The aforementioned model in 1.30 is fitted thanks to the partial likelihood

$$\prod_{h=1}^{r_j} \frac{e^{\beta_j x_h}}{\sum_{l \in R(t(h))} e^{\beta_j x_l}} \quad (1.31)$$

1.12 An example for the Fine and Gray model

It is very vital to assess the significance of the different explanatory variables in the model of the competing risks. On top, checking of model assumptions is also very important, either when a Cox model is used or when inference has been



obtained thanks to the Fine-Gray model. Furthermore, it should be noted, that the difference between these two models is that they have different cause-specific hazard functions. Graphical inspection of cumulative incidences or cumulative hazards can give a sign the assumption of proportional hazards may be violated. An arbitrary example can help in the understanding.

Let a group of people that needs to be followed-up in order to assess the contagion and the mortality of HIV. Let split this group into two smaller groups, one with people living in poorer countries and one with people living in richer countries where the national health system is superior. Furthermore, if someone has infected from HIV, then the possible causes that can die is either Aids related, or liver related or even non-natural. Of course, as it has already been mentioned, if someone dies for example from a liver failure, then he/she cannot die from the other causes.

The location of the group plays significant role, since it affects the incidence of the possible causes. In poorer countries, where the living conditions are worse, some causes of death may give higher hazards. However, if we fit the Cox model and model cause-specific hazards, we cannot infer what are going to be the consequences of the location. The reason is that the cumulative incidence of every given cause of death is determined by the risk of all potential events of interest. The solution can be given, with the Fine and Gray model where the possible effects of the location can be fitted. If the scope is to evaluate the risk factors for various outcomes, then the Cox proportional hazard model can be also used and give accurate results.

For the example above, the model for the sub-distribution hazard function for an individual in the i th location that dies from the j th cause is $\lambda_{ij}(t) = \lambda_{0j}(t)e^{(\beta_j x_i)}$, where $x_i = 1$ if the location is a rich country and $x_i = 0$ otherwise. In the model, β_j can be defined as the sub-hazard ratio between the locations for a specific cause of death. With the help of the Cox model or the Fine Gray model and with the usefulness of confidence intervals and p-values, inference can be obtained, depicting or not significance of the location to the cause of the outcome. Also, even if it is usual to fit n models for n different events of interest, the fitting of one cause-specific hazard model is a good alternative. In order to use the latter approach, the subject must have n records for the n causes and an index that clarifies in which event of interest the corresponding subject belongs

The basic ways that someone can fit cause-specific hazard models are two. The first is stratification according to event type. This method yields equal estimates to those obtained by fitting various proportional hazards models for various failure causes. The alternative is to consider the event type as a covariate. As a



result, there are proportional baseline hazards between the event types that are in scope. Of course, there are other ways such as the use of interaction terms in the covariates. With the last method the effect of the covariate can be calculated on different event types and someone can check the assumption if there is a statistically significant effect. Moreover, the model for the cumulative incidence of death from cause j for the subject i is

$$F_{ij}(t) = 1 - e^{-e^{\beta_j x_i} L_{0j}(t)}, \quad (1.32)$$

where $L_{0j}(t)$ is the baseline cumulative sub-distribution hazard function. Like the explanation above, if $x_i = 1$ then we assume that the location is a rich country and $x_i = 0$ otherwise. There are two different procedures that someone can follow in order to estimate $F_{ij}(t)$ (*Giorgos Bakoyannis and Giota Touloumi, 2012*). The first procedure has to do with the replacement of the event time for someone that has died from a competing cause different from j . In that occasion, the fixed censoring time takes the place of the event time.

The alternative is the classic one, to use the data that you have in order to fit the Cox proportional hazard model. Moreover, it should be mentioned that fixed censoring times may be used because the actual censoring time of an individual until he/she dies from the competing event is not known. Different procedures have been proposed like the one based on multiple imputations or the one with inverse probability of censoring weighting (*Fine and Gray*). In both procedures, it is possible that dependence of the censoring distribution on a set of values will be in place. In general, in the majority of the occasions, random right censoring is the ultimate step.

As we have already mentioned, it is crucial to check the assumption of proportional hazards. Depending on the example, if in the beginning of the study the incidence of death for the one location exceeds that in the other location and at later times this has been reversed, then there is an indication that the assumption may not be valid. This indication can be found from graphical curves or with tests. This situation may happen because of time-dependent covariates. For example, the covariate for the variable of location can be considered as time-dependent and as a result, its value can change through time. The estimation of the effect of time-dependent covariates can be accomplished with the use of the Fine and Gray model. Although it is necessary to understand the problem and know the time-dependent value of the covariate throughout the follow-up period.

Moreover, it should be noted that there are two groups of time-varying covariates. The one group is covariates that are external, namely their value is changing



in any case and is not depending on individual status. One characteristic example is age. These covariates can be used in the in both the cause-specific hazard model and the Fine and Gray model. On the other side, the other group contains the internal variables, namely variables that are dependent on the survival of the individual, such as the blood pressure throughout the follow-up period. These covariates can be used in the cause-specific hazard model only. Internal time-dependent covariates cannot be used in the prediction of the CIF. For more information regarding the internal and external time-varying covariates, please refer elsewhere.

1.13 Conclusion of the first chapter

In this chapter, there was an introduction in the survival analysis and its important components. The definition of competing risks and the functions that are used were given. Competing risk analysis will be the core component of the thesis regarding the HIV data with missing event types. In our problem there is a misclassification of the event types since there are a lot of unreported deaths that have been wrongly classified as disengagements from care. Our approach, as we have said in the introduction above, will be to transform the misclassification data to missing data (IeDEA dataset) and then proceed with the competing risk analysis and estimate the event types, where the two event types will be the death from HIV and the disengagement from care. In the next chapter our main interest will be the definition of the missing values and the misclassification.





Chapter 2

Dealing with misclassification and missing event types in competing risk settings

2.1 General idea and description of the problem and a Bayesian definition

The data that will be used have been extracted by the *East African International Databases to Evaluate AIDS (IeDEA-EA)*, which is one of the seven regional data centers that are funded by the United States National Institute of Health to gain valid data for the HIV disease. This region consists of Kenya, Tanzania and Uganda. According to the article (*Yiannoutsos et al. ,2008*), that was about a related survey, adult HIV-positive patients are recorded by health workers that have obtained contact information in order to trace the patients in the case of disengagement. In the article (*Yiannoutsos et al. ,2008*), all subjects have initiated the treatment from 01/01/2010 and after. In the dataset that will be used in the next chapters, all the subjects have initiated ART treatment between 12/02/2004 and 20/4/2017. In the aforementioned article (*Yiannoutsos et al. ,2008*), an initiative of tracing HIV-positive patients had been created and included patients that received either a simplistic ART or a combination ART (cART).

One of the primary objectives of collecting HIV data from these regions is to identify potential risk factors for death and disengagement from care, and to pro-



vide specific prognosis and prediction estimates for these two outcomes. Unbiased estimation is crucial for health authorities in order to effectively target interventions and identify patient groups that are at a higher risk of disengagement. These patient groups may include individuals at an early -the healthier individuals are less motivated to stick to HIV care- stage of the disease or those facing challenges accessing healthcare due to the distance of surveillance health centers from their homes. In the dataset to be used, a person is considered to have a gap in care when she/he fails to visit the health center for more than two months after the scheduled visit date. It is important to note that the first date a person is classified as a dropout is exactly two months after the next scheduled visit.

Furthermore, as mentioned in the first chapter, a significant challenge in analyzing competing risks data, which is the focus of interest here, is that one cannot simultaneously estimate the cause-specific hazard function and the cumulative incidence hazard function using a single approach. Additionally, in this region, the lack of resources and awareness about the disease presents a major barrier. This not only leads to statistical misjudgments but also affects the overall infrastructure and validity of the results.

As it has already been mentioned in the conclusion of the first chapter, regarding motivating IeDEA dataset, a lot of deaths are incorrectly classified as disengagements from care. As a result, a misclassification issue is emerging. Misclassification arises when a subject is incorrectly classified in another group than the group that should have been assigned. The misclassification in the HIV dataset is a consequence of death under-reporting which results in the incorrect belief that the disease is not as lethal as it seems. This is expected to result in seriously biased estimates. The assumption that the death rates of the disease are lower can have social, sanitary and financial consequences in the local communities and beyond.

Naturally, the individuals that have been classified as disengagers are either deceased, and their death is not reported to the clinic, or have actually disengaged from care. This situation has as a result the increased infectivity of the disease, since the disengagers are not followed and controlled, and the increase of mortality as a direct consequence of disengagement from care and not receiving treatment. As it has already been noted, an approach that can be followed is to transform the misclassification problem into a missing data problem and then "predict" the cause of failure of every patient. In other words, the event type from now on is considered to be unknown for the patients that had a gap in care but have not been traced from the health authorities. With that transformation, the next step is to estimate the hazard ratios of each outcome using either frequentist statistical or Bayesian approaches.



The existence of missing data is expected to negatively impact future analyses. First of all, the estimators will be biased if we just use the complete cases (i.e., without missing data) in the analysis. Although, it is common to use computationally easier-to-handle methods, even if they tend to be biased and have lower statistical efficiency or need strong assumptions to follow.

The well-known and significant approaches to deal with missing data are the following. One approach is to exclude the missing data completely from the analysis and just estimate whatever you need with the complete data (complete case analysis). Furthermore, imputation methods can be used such as the mean or median imputation. Whether the missing values have been omitted, problems like the reduction of the statistical power of the model or incorrect conclusions can emerge from the analysis and our new dataset will be considered as biased. On the other side, the approach that imputes data using the mean can be characterized as a bad choice, since it does not take into account, the relationships and the correlations between variables and the standard errors are usually underestimated. Another major approach is to impute the missing data using a modelling approach with values that have been estimated from models based on the complete data. A very popular approach is MICE (*Multivariate imputation by chained equations*) which can be applied in situations where there are missing values in multiple variables. Moreover, the use of the predicted probabilities of a model that have been fitted thanks to the complete cases is very common and can give precise results if the structure of the model is correct. Last but not least, missing data can be analyzed with inverse probability weighting (IPW) approaches. The IPW approaches are based on the logical notion of removing selection bias caused by missing data by establishing a pseudo-population of weighted copies of the complete cases. Nevertheless, based on the method and pattern of the missing data, various weighting approaches should be considered.

Dealing with missing values cannot only be done using approaches from frequentist statistics. As we will see in chapter four, and it has already been mentioned, the Bayesian approach will be used and can be very helpful in the problem with the missing values. In general, the main difference between the classical approach and the Bayesian is that in the former for the estimation of a parameter, the likelihood function $f(y|\theta)$ is used and in the latter for the estimation, the posterior distribution $f(\theta|y)$ is used. In Bayesian statistics, the assumption is that the parameters that need to be estimated are not fixed but random variables. The posterior distribution will be constructed using both the prior distribution and the likelihood of the dataset, namely



$$f(\theta|y) \sim f(y|\theta)f(\theta), \quad (2.1)$$

where $f(y|\theta)$ is the likelihood and $f(\theta)$ is the prior distribution. The likelihood can be created easily from the data that are known, as well as the prior distribution that can be created by extracting information from the observed data or a previous analysis on the same topic or even an insight from an oral conversation. Naturally, some generic non-informative prior distributions can be used and give precise results. The prior distribution can be regarded as the information that is already known for the parameter. An example regarding the prior distribution can be the relative proportion of voters that will vote for a particular political party in the future or the information that a specific subset of people tends to have the disease due to their genealogical tree. Long story short, Bayesian statistics views uncertainty as a subjective belief and incorporates prior knowledge and data to update this belief. It treats unknown parameters as random variables and provides a posterior distribution that combines prior information and observed data.

In the Bayesian framework, missing data is treated as a random variable and can be estimated using conditional distributions based on the available data. The same principle applies to the estimation of future predictions. Bayesian methods can be applied to estimate either the response variable or covariates in a model. Various approaches are available within the Bayesian framework, with two well-known and significant ones being Gibbs sampling and the Metropolis-Hastings algorithm. In this paragraph, we will primarily focus on the Gibbs sampling procedure, which is a specific type of Markov chain Monte Carlo (MCMC) method. Gibbs sampling leverages the fact that a joint distribution can be fully specified by a set of conditional distributions, known as full conditional distributions. It should be noted that while the predictive posterior distribution in MCMC is

$$f(Y, \theta|y) \sim f(Y|\theta)f(y|\theta)\pi(\theta), \quad (2.2)$$

with Y be a new parameter, the corresponding predictive distribution for the Gibbs sampler only needs the first part of the 2.2, namely

$$f(Y, \theta|y) \sim f(Y|\theta), \quad (2.3)$$

Gibbs sampling is an iterative method that samples from the conditional probability distribution of the given variables, based on the initial values assigned to these variables. It is a random walk procedure where a proposal is drawn for each new value of the parameter to be estimated, using the conditional posterior probability distribution as the proposal distribution. Notably, the proposed move is



always accepted. However, it is important to accurately specify the conditional probability distributions for the approach to be effective. One limitation of Gibbs sampling is its potential slowness due to parameter correlation. This occurs because the random walk can only take diagonal steps, which can hinder exploration of the parameter space.

The Multiple Imputation by Chained Equations (MICE) methodology, implemented in the R programming language, leverages Gibbs sampling as its core component. MICE employs Gibbs sampling to create multiple imputed datasets, allowing for proper assessment of the uncertainty associated with missing values. In each iteration, values are generated stepwise (e.g., in step 'k') using the fully conditional distributions estimated in the previous step. It is crucial to correctly specify the priors used in this process. As mentioned earlier, missing data is treated as unknown parameters, and the Gibbs sampler estimates these parameters by replacing them with the missing values.

2.2 The two possible approaches to tackle the issue

A lot of individuals that have been tested for the disease and received antiretroviral treatment (ART) had a gap in care and classified as disengagements. ART is considered to be very efficacious in the treatment of HIV since according to early studies it reduces mortality. Therefore, it is very important to tackle the issue of death under-reporting. There were two possible alternatives that should have been followed in order to solve the problem of under-reported deaths.

The one alternative was to search for all the individuals that had a gap in care. This procedure demanded the researchers to investigate, even door to door, all the individuals that are considered as disengagers. This surveillance approach also included communication with family and friends, telephone contacts and other procedures. This alternative had to be limited to a relatively small number of individuals since it was too expensive both financially and in terms of human hours and time. The solution to this problem is a double-sampling design where only a small number of patients that have been classified as disengagers are being outreached.

In our problem, the double-sampling procedure consists of two distinct steps.



The first step involves selecting the sample of individuals who have disengaged from the total population. However, it is important to note that this step may be imperfect if many individuals in the sample would not be included in the final analysis. This sample, which contains all the disengagements in our dataset, is assumed to include unreported deaths. Subsequently, a second sampling takes place within the initial sample that has already been collected. The goal of this second sampling is to identify and locate individuals who have a gap in care. This step relies on the sample of patients who have been lost from treatment and subsequently traced. Using this sample, we can estimate the probabilities of the various outcomes for those individuals who have not been found and validate the unreported deaths. By employing this double-sampling procedure, we aim to capture both the true disengagers and the unreported deaths, as well as estimate the outcomes for individuals who could not be located. The last sample contains the patients where tracing was held and these patients are considered as outreached. This approach helps to address the challenges associated with missing or incomplete data in order to obtain more accurate and comprehensive results. This step is characterized as an Expensive gold standard event type ascertainment (*Bakoyannis et al., 2020*). A detailed analysis of the aforementioned approach is portrayed in Figure 2.1.



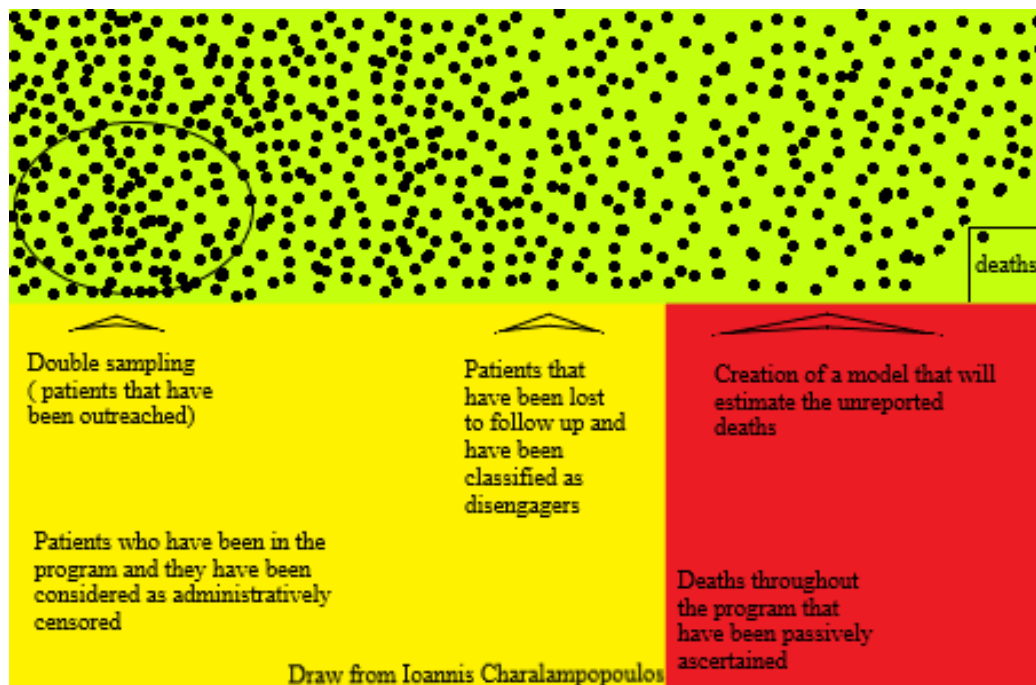


Figure 2.1: The approach that is followed regarding the estimation of the unreported deaths in the subsample of the patients that have a gap in care.

The double-sampling approach is used in order to reduce the bias that would be induced by just using the fully observed data. There may be reasons that influence the patients that have gap in care and do not appear in the group of patients that did not leave the treatment. Moreover, the statistical efficiency of the estimates, namely the precision of the estimates, would be lower, since we would discard a substantial amount of observations (those with missing event types). Although, when the analysis takes place, even if our interest is focused on the patients that have a gap in care, patients that are until the end of treatment and have not experienced the outcome of interest, namely they are considered as right censorings, will not be omitted because if they are excluded, bias will increase.

Furthermore, as it will be discussed later, predictions and inference for an individual's risk is not possible. Also, patients that have a gap in care and patients that have observed status have differences in survival and as a result, if this approach is not followed then the overall mortality of the disease could be underestimated. With the underestimation of disease mortality, the validity of these programs will

suffer a great hit and biased estimates will lead to invalid conclusions.

2.3 Approaches for the estimation of the mortality of the disease and the probability of a cause of failure to take place

Different approaches for estimating the mortality of the disease have been proposed, some of them have the Kaplan-Meier estimation as the center of interest. One method is the classic Kaplan-Meier estimate, which is calculated using the deaths that have been recorded. As it is already known, this method is naive since the results will be biased. Another method is to also use Kaplan-Meier but instead of the observed values, use all the deaths that have been either obtained with a record or traced after gap in care. This method is expensive and no statistical method in order to estimate the lost from treatment deaths is used.

The third approach is to create a weighted average of the Kaplan-Meier estimate between the group of disengagers and the group with the observed status. Although this method is considered to be better than the aforementioned ones, since the weights give more to the point estimations, it is also considered to be biased. The last method (*Frangakis and Rubin, 2001*) proposes to use the weighted average hazard of death between the two groups and not the Kaplan-Meier survival estimate. Using this approach, the survival estimates were shown to be consistent.

For the sample of the “disengagers” from care, as has already been mentioned, there is no evidence if they are dead or alive. A binary logistic model was fitted in order to calculate the probability of an individual to be deceased or survived and just left the survey. The information that has been gained from the second sampling regarding the unreported deaths can be used in the logistic models in chapter three and chapter four, where the latter will be considered as Bayesian. The given information can be used as the prior probability for the model. Furthermore, based on previous investigations (*Bakoyannis et al., 2020*), the goodness of fit of the model is not considered to be robust concerning the model’s assumptions.

As it has already been noted, the knowledge regarding the cause of failure in the problem is incompletely observed. The outcome is not known for all the sub-



jects of the dataset. A huge variety of approaches have been utilized to calculate the cause of failures such as the EM algorithm for estimation in the method of piecewise-constant hazards competing risks (*Craiu and Duchesne, 2004*) and the method with the augmented inverse propensity estimators (AIPW) for the regression coefficients (*Gao and Tsiatis, 2005*).

2.4 Augmented Inverse Propensity Weighted estimators

Even if this approach will not be detailed discussed in the thesis, it could be used in the comparison with the MPPL (Maximum pseudo-partial likelihood estimator), which will be our final approach. This estimator is a relatively new approach (*Robins and Rotnitzky, 1992*) and is used for estimation when the dataset includes a lot of incompletely observed data. *Gao and Tsiatis (2005)* created augmented inverse probability weighting estimators for every regression coefficient of the semi parametric linear transformation models and as a result the probability of missingness as well as the probability of every outcome of interest can be estimated. Further work from *Hyun et al. (2012)* was made and as a result the AIPW approach could be applied in proportional cause-specific hazards models. These AIPW estimators have the double-robustness property and are more effective than the simple inverse probability weighting estimators. Even if one of the parametric models for the probability of missingness and the cause of failure is wrongly defined, the double-robustness property guarantees consistency. Moreover, it should be mentioned that in the AIPW estimator double robustness property is considered to be followed. Double robustness means that the AIPW approach remains unbiased and consistent even if one of the parametric models that is used, is misspecified. This approach also strengthens its position due to the fact that it gives higher statistical efficiency in comparison to the simple inverse probability estimators. Although, big drawback is the fact that if all the parametric models, that are used to deal with missingness, are wrongly specified then the estimates will be biased.



2.5 Multiple Imputation approach

Another approach, that will be used for comparison with MPPL, is the Multiple Imputation (MI) approach. In general, multiple imputation is used in a dataset in order to deal with missing data. In this approach, instead of making only one imputation, multiple imputed data sets will be used. The difference with the single imputation is the fact that with multiply imputation the control and calculation of the uncertainty that derives from this approach is possible. Following this procedure, it is possible to correctly estimate the variability of the estimates. The procedure that is used is split into two parts. In the first step, the missing values are replaced in the dataset through imputation p times and then in the second step the imputed data sets are analysed and a combined result is used, through simple calculations like the mean.

In order to obtain valid results, it is important that assumptions must be met, such that the imputation model needs to be correctly specified. If this assumption does not hold, then a transformation in the data is needed. Moreover, the computational speed of this procedure is not a strong point, because this analysis must be repeated multiple times. Furthermore, it should be noted that also in this procedure, the analysis will be biased if there are not enough variables that can help in the prediction of missing values. Last but not least, practical implications can reduce the precision of the procedure. The importance of the variable can be taken into consideration whether or not it carries missing data.

In this thesis, there will be no further description of these methods because of the computational difficulty that arises. Our main approach for the frequentist part of the thesis will be to use the maximum pseudo-partial-likelihood estimation (MPPL) method for the semi-parametric proportional hazards model. Thanks to this approach, supplementary covariates will be used for dealing with the missing cause of failure problem and the implementation is considered to be more straightforward.



2.6 Important notation for the approach in scope

For our problem, we should give the appropriate notation. One of the key assumptions that the model should follow is the missing at random assumption (MAR). This assumption tells us that there is a relationship between the tendency that there are missing values and the actual observed data, but not with the actual missing cause of failure. Given that C is the true outcome that an individual will have, C^* is the possible false cause of failure that has been observed and R_i is an indicator that examines whether or not the cause of failure is known ($R_i = 0$ indicates that the cause of failure has not been observed and $R_i = 1$ otherwise), then the probability that the outcome j has appeared to the patient i is

$$P(C_i = j | R_i = 1, C_i^* > 0, W_i) = P(C_i = j | R_i = 0, C_i^* > 0, W_i) \quad (2.4)$$

$$= P(C_i = j | C_i^* > 0, W_i) = \pi_j(W_i, \kappa_0), \quad (2.5)$$

where W_i is a set of auxiliary covariates of interest that are probably related to the probability that a value is missing. The missing at random assumption (MAR) tells us that the probability that was estimated by the model that has been used in the double-sampling sample is the same to the probability in the sample where the cause of failure has not been observed, namely $R_i = 0$. Furthermore, assuming that X is the failure or the right censoring time and Δ_{ij} an indicator that records if for subject i the cause of failure is the type j ,

$$N_{ij}(t) = I(X_i \leq t, \Delta_{ij} = 1), \quad (2.6)$$

The aforementioned formula (2.6) is the cause-specific counting process, namely it is an alternative approach that allows the separate modeling of each competing event. Instead of treating all competing events as censored, the cause-specific counting process assigns individuals to specific risk sets depending on the event type they are at risk at each time point. The formula 2.6 allows us to count the number of subjects that have expressed the outcome j until time t . Moreover the model that will be fitted for the estimation of the outcome of interest is a logistic model, so $\pi_{ij}(W_i, \kappa_0)$ is a binary logistic model if two are the possible outcomes, as it can be probably considered in the thesis, where κ_0 is a finite-dimensional vector that includes the parameters of the logistic model (*Bakoyannis et al., 2020*). If more causes of failure were in place, then the model would be a multinomial logit



model with a generalized logit link function. In our problem, since the possible competing risks are two, the aforementioned probabilities of the relationship 2.6 is in place

$$\pi_2(W_i, \kappa_0) = 1 - \pi_1(W_i, \kappa_0). \quad (2.7)$$

Moreover, it should be noted that the logistic model (please refer to formula 1.30) may be misspecified. If it happens, two are the steps to deal with model misspecification. First of all, the initial step is to create a goodness of fit evaluation in order to see whether there is evidence for lack of fit in the model. If there is evidence that there is lack of fit, then approaches such as logarithmic or B-spline terms can be used. The main goal is to estimate the corresponding estimators for the model in scope. If the cause of failure was known for the total number of subjects, then the vector of the estimates of the proportional hazards model (1.30) could be estimated with the maximization of the partial likelihood

$$pl_n(\beta) = \sum_{j=1}^k \sum_{i=1}^n \int_0^{\tau} [\beta_j^T Z_i - \log[\sum_{l=1}^n Y_l(t) e^{\beta_j^T Z_l}]] dN_{ij}(t) = \sum_{j=1}^k pl_{n,j}(\beta_j). \quad (2.8)$$

However, a different approach needs to be followed, since not all the possible outcomes are observed for the subjects and the aforementioned formula is not anymore applicable. As it has been mentioned, the disengagements have an unknown cause of failure, so the partial likelihood that will be estimated is based on the observed data, namely the subjects where the cause of failure is known. Moreover, regarding the cause-specific counting process, a new formula will be used and the difference with the formula 2.6 is that instead of $N_{ij}(t)$, $\hat{N}_{ij}(t)$ is in place, where

$$\hat{N}_{ij}(t; \kappa_0) = [R_i \Delta_{ij} + (1 - R_i) \pi_j(W_i, \kappa_0)] N_i(t). \quad (2.9)$$

Looking at the aforementioned formula, Δ_{ij} is the event indicator for subject i and it is assumed that every subject is known that will have an event of interest. The main difference between $\hat{N}_{ij}(t)$ and $N_{ij}(t)$ is that in the latter the occurrence of the one cause of failure has as a result the non-occurrence of the other cause of failures. On the other hand, with the formula 2.9 it is assumed that the probability of the one cause of failure to occur does not necessarily means that the other cause of failure has zero chances to take place. In general terms, it should be highlighted that the expression inside the brackets of formula 2.9 tells us that if the event of interest is not known, then the expression is



$$\hat{N}_{ij}(t; \kappa_0) = \pi_j(W_i; \kappa_0)N_i(t), R_i = 0. \quad (2.10)$$

2.7 A toy example

The primary objective of our study is to implement a double-sampling approach within the disengagers' sample. This double-sampling procedure involves the inclusion of disengaged patients who have been successfully traced, enabling us to ascertain their real status. By utilizing the data from these individuals, we can construct a logistic regression model within this sample. This model facilitates the estimation of hazard ratios for the covariates and help us to predict the number of deaths and estimate the coefficients for the remaining disengagers. Furthermore, the accuracy of the model can be assessed using either classical or Bayesian statistical methods.

Very important is to understand how formula 2.9 works. Let's create a toy example for a better understanding. The numbers in the example are random, focusing on the explanation of the mathematical formula and not on the explanation of a specific problem. Let 's suppose that $X = \min(T, U)$ is the time from the beginning of the survey until the event of interest or the censoring, whatever comes first. As has already been mentioned, Δ_i is an indicator, in which i is the event of interest, where 0 indicates censorship and 1 that an event has taken place, while R is an index that indicates whether the event of interest is known or not. Furthermore, suppose that the interest is to model the indicator Δ_i and let $\hat{\Delta}_i$ be the probability given the covariates, i.e.,

$$P(\Delta_i = 1 | \Delta = 1, W), \quad (2.11)$$

where W is the vector that contains the covariates of the model.

Suppose that there are six patients and the explanation of the formula 2.9 is depicted in Table 2.1 with arbitrary examples.



I_d	X	Δ	R	Δ_1	Δ_2	$\hat{\Delta}_1$
1	4.2	0	1	0	0	0
2	2.4	1	0	?	?	0.35
3	3.5	1	1	1	0	1
4	2.9	1	1	0	1	0
5	5.6	0	1	0	0	0
6	7.1	1	0	?	?	0.5

Table 2.1: A toy example that depicts the possible values that the indexes Δ and R can have as well as the calculation of Δ_i .

However, it is known that $\Delta = \Delta_1 + \Delta_2$, so for example for the second patient, it is certain that either Δ_1 or Δ_2 will be one since one of the event of interest will take place. Looking at the first and fifth patient, it is observed that there is no event of interest since the event is censored, namely the patient did not express none of the possible outcomes of interest. As a result, the value in the last column is zero as well as the values of $\hat{\Delta}_2$. In the third example, it is supposed that the patient has expressed the first event of interest, since $\Delta_1 = 1$. In both occasions, it is obvious that $R = 1$ and $\Delta_1 = \hat{\Delta}_1$.

The same result will happen regarding $\hat{\Delta}_2$ for the fourth patient since it is known that this subject has expressed the second outcome. The outcome for the second patient is not known and as a result $R = 0$. In this occasion, $\hat{\Delta}_1$ should be calculated. The value in the example is a random probability since the example is arbitrary. The value will be calculated based on the covariates that the example would have.

2.8 Important functions for the implementation of MPPLE

If instead of the unknown κ_0 parameters, a $\hat{\kappa}_n$ estimator is used in the partial-likelihood, then the estimation of these unknown parameters is possible. With a $\hat{\kappa}_n$ estimator a pseudo-partial-likelihood can be constructed. Due to the fact that a consistent estimator is used and not observed values, the partial likelihood is called pseudo. The first step is to maximize this partial likelihood, which is constructed thanks to the subjects where the cause of failure is accurately observed. The second step is to calculate the estimators $\hat{\beta}_{n,j}$ for the subjects of the research. To do that, one must solve the following equation

$$G_{n,j}(\beta_j, \hat{\kappa}_n) = 0, \quad (2.12)$$

where

$$G_{n,j}(\beta_j, \hat{\kappa}_n) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau [Z_i - E_n(t; \beta_j)] d\hat{N}_{ij}(t; \hat{\kappa}_n), j = 1, \dots, k \quad (2.13)$$

and

$$E_n(t, \beta_j) = \frac{\sum_{i=1}^n Z_i Y_i(t) e^{\beta_j^T Z_i}}{\sum_{i=1}^n Y_i(t) e^{\beta_j^T Z_i}} \quad (2.14)$$

The function in 2.13 is known as the Cox score function or pseudo-score function. We follow this procedure in order to avoid simulating the cause of failure and use prior probabilities for the possible outcomes of the subject and as a result increase the statistical efficiency of the estimators. Following that approach, it is possible to estimate the necessary estimators for the proportional hazards model. An important first step is to assess the parametric assumption of the model $\pi(W_i, \kappa_0)$.

This can be achieved using the cumulative residual processes

$$L_j(t; \hat{\kappa}_n) = E[R_i[N_{ij}(t) - \pi_j(W_i, \kappa_0)N_i(t)]], \quad (2.15)$$

which can be estimated thanks to

$$\frac{1}{n} \sum_{i=1}^n R_i[N_{ij}(t) - \pi_j(W_i, \hat{\kappa}_n)N_i(t)], j = 1, \dots, k \quad (2.16)$$



for $t \in [0, \tau]$

In order to confirm that the parametric model is correctly specified, the expected value of the formula 2.16 should be equal to zero, namely

$$E[L_j(t; \kappa_0)] = 0, \text{ for all } t \in [0, \tau] \quad (2.17)$$

In addition to the use of the formula 2.16, Pan and Lin (2005) created a rigorous goodness of fit that strengthens the parametric assumption of the model. This goodness of fit can be also evaluated graphically, where the $(1 - \alpha)$ confidence band around the residual process plays an important role.

If the confidence band fully covers the vertical line $y = 0$, then there is no evidence for lack of fit. Otherwise, the null hypothesis of a good model fit is rejected. In the latter case, one can use flexible terms such as B-splines to alleviate the consequences of a poor model fit. According to the article of Bakoyannis et al. (2020, p.22), even if some competing events of patients are not accurately predicted, the estimators can be said that have a good performance.

Last but not least, in order to estimate the cumulative baseline cause-specific hazard for the event of interest for the i subject, one can use the Breslow-type estimator, namely

$$\hat{\Lambda}_{n,j}(t) = \int_0^t \frac{\sum_{i=1}^n d\hat{N}_{ij}(s; \hat{\kappa}_n)}{\sum_{i=1}^n Y_i(s) e^{(\hat{\beta}_{n,j}^T Z_i)}}, j = 1, \dots, K, t \in [0, \tau]. \quad (2.18)$$

It should be noted that the MPPL estimator is consistent and asymptotically normal. In the formula 2.18, $\hat{\kappa}_n$ is the consistent estimator. An estimator is considered to be consistent when the estimated value converges to the true value of the parameter with the increase of the sample size to infinity. So, the estimator is going to be representative of the actual value. Moreover, when an estimator is asymptotically normal, its distribution converges to a normal distribution with the increase of the sample size.

Furthermore, the follow-up period should not be considered as infinite, namely, the interval is $[0, \tau]$, with $\tau < \infty$. Moreover, whether the model in scope, $\pi_j(W_i, \kappa_0)$, is a properly described binary or multinomial logit model, and the model's parameters are estimated using maximum likelihood, then automatically the inverse g of the link function for the parametric cause of failure probability model $\pi_j(W_i, \gamma_0)$, $j = 1, \dots, k$, has a continuous derivative \dot{g} with respect to on compact sets (Bakoyannis et al., 2020).



2.9 The constructed models of the frequentist's statistics

The core of the analysis is to set up the model. The model will be created at first with the classical statistics procedure and then with the Bayesian. First of all, it is very important to create a general model that can be used specifically for our problem. In our dataset there are two causes of failure, death and disengagement and the Cox proportional hazards model can be fitted first at the sub-sample of the patients that have a gap in care and have been outreached. In chapter four, the same procedure in the Bayesian framework will be applied to the dataset. As we will see in the next chapter, three frequentist procedures will be followed.

The first approach will be a simple Cox proportional hazards model for the complete cases. One of the variables of interest is the time until the cause of failure emerges so Cox regression is a better choice than a simple logistic model. It should be noted that the Cox proportional hazards model relies on the proportional hazards assumption, namely the survival curves of the different groups should be proportional over time and should not overlap with each other.

The other two approaches that will be presented in the next chapter are the following. In the first one, the procedure of Multiple imputation is taking place. The first step of this approach is to impute the missing data using the multiple imputation approach and specifically logistic regressions will be fitted 100 times, where the missing values will be simulated from the predicted probabilities of the logistic model. As it has been said, this procedure does not depend on a single model run but on multiple iterative applications of the method. The second step is to create a Cox proportional hazards model as the one in the previous approach with the difference that in this occasion there is an estimation of the missing data thanks to the imputation. Moreover, thanks to Bootstrap, the standard errors can be precisely estimated.

The last one, and hopefully the best one in the third chapter, will be the semi-parametric proportional hazards model for competing risks data with the missing cause of failure that uses the maximum pseudo-partial-likelihood estimation (MPPLE) method (*Semiparametric regression and risk prediction with competing risks data under missing cause of failure. (Bakoyannis et. al.,2020)*). In chapter four the cause-specific hazard model with Weibull baseline hazard function will be considered. This model will be considered under the Bayesian framework. As



it has already been noted, one of the main approaches in scope is the MPPL. In order to prove that this estimator is the best, it is important to compare it with the other approaches such as the complete case analysis and the multiple estimator (MI). Moreover, the standard errors of these estimations should be calculated in order to evaluate the relative efficiencies of the different methods. For multiple imputation and MPPL approaches, bootstrap will be used.

2.10 Conclusion of the second chapter

In the aforementioned chapter, a review of some important papers regarding competing risks and the approaches that can be used for the estimation of the corresponding probabilities of these competing risks took place. The two most important papers that were the core of this chapter are the ones by (*Bakoyannis et al., 2020*) and (*Yannoutsos et al., 2008*). In the next chapter, a detailed analysis of the dataset will take place as well as the check of the relationships between the variables and the creation of the aforementioned models in the frequentist framework.



Chapter 3

Description of the disease, the dataset and creation of the models in scope

3.1 Description of the disease

The two Lentivirus species that infect humans are the human immunodeficiency viruses (HIV). Lentiviruses are a subclass of retroviruses. They eventually lead to acquired immunodeficiency syndrome (AIDS), a disorder in which the immune system is gradually weakened, allowing malignancies and life-threatening opportunistic infections to proliferate. It is commonly transmitted via the sexual act, but can also be transmitted via blood or sharing needles. It was first observed in 1981 in the United States and it is considered a mutation of the simian immunodeficiency virus (SIV) that infects chimpanzees. It has been widely transmitted in the African continent and this is the main reason that a lot of research has taken place in this region.

The principal life-threatening consequence for a patient that has been infected with this virus is the progressive weakness of the immune system. As a result, the immune system of the patient is more vulnerable against infections and some types of cancer that people with a healthy immune system can more easily fight off. At



the beginning of the disease, the patient is usually asymptomatic and as a result she/he is not aware that she/he can transmit the disease. Depending on the immune system, years can pass by but the result of the weakness of the immune system is certain. In regions like Africa, where the protection and awareness of HIV is lower than in richer countries, this plays a significant role in the transmission of the disease. The main symptoms as the disease progresses are swollen lymph nodes, weight loss, fever, diarrhea and cough.

HIV infection eventually leads to AIDS. In this phase there is a progressive reduction of the count of CD4 cells. A subset of white blood cells is CD4 cells. They are also referred to as "helper T cells" or CD4 T lymphocytes. That's because they aid in the prevention of infection by supporting the immune system to eliminate bacteria, viruses and other potentially harmful pathogens. HIV attaches itself to the CD4 molecule on the surface of helper T-cells and replicates within them. This causes the destruction of CD4+ T-cells, resulting in a steady decline in this T-cell population. An indication that a person has AIDS, is the count of CD4 cells to be lower than 200 cells/mm³.

3.2 Description of the dataset

As it has already been referred above, the dataset that will be used for the description and solution of the problem has been given with the authorization of the *IeDEA* (i.e., *international epidemiology Databases to Evaluate AIDS*). First of all, there are 71 already fixed variables in the dataset that help the health authorities to get a better understanding of the disease. For example, the variable *ptidno* helps the research to uniquely identify each patient, since a distinct number is given to each one of them. Moreover, there are variables that record the dates of viral load measurements (*VLdate*). A patient receives a viral load test in order the level of the human immunodeficiency virus (HIV) in the blood to be checked, namely the amount of genetic material (RNA) of HIV in patient's blood, and each patient receives numerous tests, 20 in this survey, since viral load can be different in every visit and it depends on the level of the disease. If the condition of a patient is worse than the last visit, this is expected to be visible in the next viral load test.

Since, there are a lot of gaps in care, the vast majority of the values of the variables is not applicable. Some patients are not attending all the necessary visits and this has as a result, a possible comparison not to be possible. The same happens for the next variable, *VLsuppress*. The variable *VLsuppress* is used in order to test whether the viral load in each patient is greater than 1000 copies/ml or not during



the ART treatment. It should be noted that the first measurement is considered to be the baseline level and if the viral load of a patient is increasing and surpass the threshold of 1000 copies/ μ l then it is a sign that the disease is getting worse. Furthermore, there are variables that record the day of enrollment (*i.e.*, *enroldate*), the date of death (*i.e.*, *death-d*), or the date of the outreach encounter (*i.e.*, *ordate*), the program that the patient is enrolled (*i.e.*, *PROGRAM*) and the average number of visits per week (*i.e.*, *aveVisPerWk*). Although, all these variables will not be used due to the fact that a lot of entries are missing.

Starting with simple descriptive statistics about the variables of the dataset, the initial dataset includes 79560 subjects that receive ART treatment. From the total number of patients, 52136 subjects are females, from which 4937 were pregnant at enrollment, and 27424 are males, while all the patients are from Kenya. All the patients have been considered as subjects of the *AMPATH* program (*Academic Model for the Prevention and Treatment of HIV/AIDS*). This program is very important in the battle against HIV/AIDS, since it is a program that traces patients that have a gap in care. The mean age of the subjects is 36.8 years, since the majority of the subjects are young adults between 20 and 40 years old, while the 95% of them are older than 21.

However, because of the fact that a lot of patients have incomplete data regarding their enrollment in the program, the dataset will be reduced to the ones that give a better insight for our analysis. Patients that did not give information regarding the CD4 cells, their gender, their status regarding the location, their age and whether their HIV status is disclosed or not, namely initial information that is crucial for the understanding, have been omitted. The new dataset concludes data from 21307 patients. Please refer to the histogram 3.1 for the variance of the age of the subjects in the final dataset.



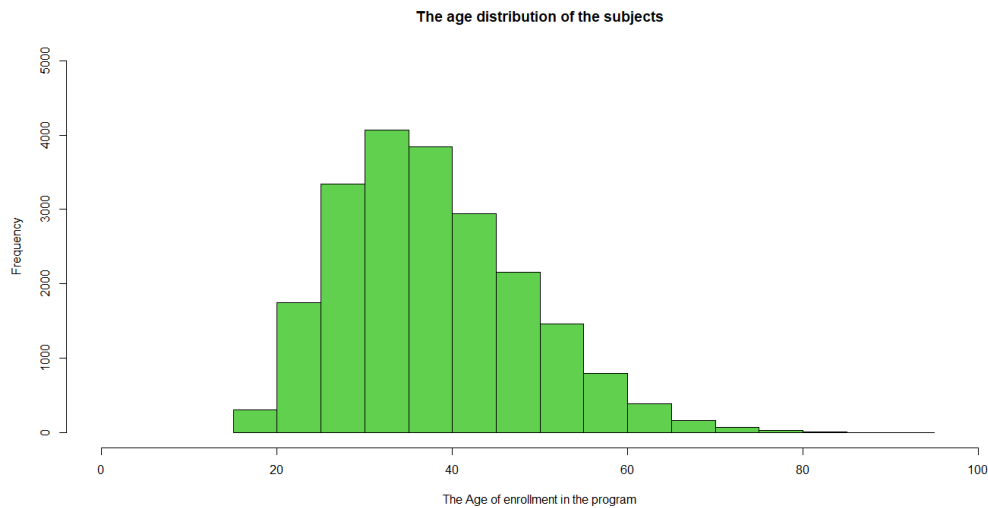


Figure 3.1: The age distribution of the subjects in the final dataset that are in the dataset, depicting that the majority of the patients are between 20 and 40 years old.

Looking at the dataset, regarding the place that the visits are taking place, 57.3% of the total population of the patients is receiving the treatment in a district hospital, 21.7% is visiting a health center and 21% a regional, provincial or university hospital for the regular visit. Regarding the variable “rural” that depicts the location of the health center, 62.8% of the patients goes to an urban location, while the rest 37.2% goes to a rural. Furthermore, regarding the percentage of people that have a gap in care, 21% of the total subjects follow the regular program that has been implemented by the health authorities without any problem, while the rest 79% had at least one gap in care throughout the treatment. According to IeDEA, the corresponding variable includes gaps which immediately precede database closure.

Another interesting variable is the variable *diedwithin2mos*, which depicts whether the subject has died within two months after the next scheduled visit date recorded at his last visit. From the total number of subjects, the 8.2% of them is known that they have died while they are in care and for the rest 91.8% the result is not known, namely if they have died within this period. Also, thanks to the results regarding the possible gap in care, whether patients have died within two months after the next scheduled visit and whether their death was passively ascertained

or through outreach, it was possible to estimate the possible cause of failure for the patients of the dataset. From the total number of patients, 4473 are considered as right censorings, namely patients whose outcome may have happened after the end of the treatment or never took place. The confirmed deaths from the program are 1731 (both deaths while in care and through outreach), while the disengagements are supposed to be 15103. Moreover, it should be noted that the 40% of the disengagers were successfully outreached, namely 6089 patients. From the 6089 patients that have been traced through outreach, 1659 of them were already dead. This fact proves the fact that disengagement plays a major role in the underestimation of death in these regions and highlights the need for a solution. These deaths are nearly the 26% of the patients that were successfully outreached. As it has already been mentioned, there are patients that have been misclassified as disengagements and as a result the total number of actual disengagers is not the aforementioned one.

Censorings	Disengagements	Deaths
4473	15103	1731

Table 3.1: The table that depicts the outcome for every patient.

Furthermore, the given dataset can be split into different categories, depending on patient's tracing status. The variable *orstatus* will be used in order to define the second sub-sample that will be used for the estimations. In Table 3.2, the different outreach status are portrayed.

Label	Number of patients
Patients that have not been outreached (also contains right censorings and deaths)	13789
Patients found	6089
Patients not found	1139
Inadequate locator form	129
Distance too far	107
Tracing ongoing	16
Cannot be traced	38
Total	21307

Table 3.2: The table that depicts the possible results of the tracing procedure for the patients that have gap in care.

The number of patients that have gap in care and then been found are 6089. In this subgroup, 4430 were disengagers, while 1659 of the patients had lost their lives and wrongly were acclaimed as disengagers. This particular group is the one that will be used for the double sampling. Furthermore, the first category of Table 3.2 contains 13789 patients, where 9255 are considered as disengagers. These patients along with patients from the rest categories (except of the patients that have been found) will be characterized as missing values due to the misspecification of the outcome of interest. From them, the real number of the patients who passed away is unknown. In the rest categories of the variable orstatus, there are some patients, whose outcome in the dataset is death. These results are considered to be wrong due to the fact that the researchers cannot know whether there are deceased patients when they are not outreached.

	Total Disengagers	Double Sampling Disengagers	Missing
Values	15103	4430	10673

Table 3.3: The table that depicts the total disengagers as well as the split between the double sampling disengagers (outreached) and the missing ones.

Moreover, a very interesting variable is the one with the name “*hiddendeath*” in the dataset. This variable records the patients that have died within two months following the next scheduled visit date of the patient’s last visit. The variable splits the deaths into two categories. The death was either passively verified or ascertained via outreach. From the total population, the 91.8% of them is either not dead or no information has been received. This is the same percentage with the variable *hiddendeath*. However, useful information can be granted from the rest of the patients. From the rest, the death of 77 patients has been passively ascertained and for 1654 patients has been verified through the outreach approach.

Since there is no variable in the dataset that counts the months, namely the time, that a patient is in the treatment, a new variable was created. The variable “time” counts the months from the beginning of the treatment until the outcome of interest to be appeared. For the creation of this specific variable, the variable regarding the outcome of interest was used. In this occasion in order to calculate the time, three “outcomes” were created. As a result, depending on the group that a patient belongs, time till death or time until disengagement from treatment or time until the end of ART can be calculated. Patients that appeared the outcome of interest before the start of the treatment have been dropped out of the dataset because they considered to be out of scope for the analysis.

The median time for someone in the program that will disengage from the treatment is 3.4 months, namely 103 days. Also, the median time for someone in the program who will eventually die is 15.4 months, namely 468 days. Last but not least, the median time for the people that will stay from their enrollment till the end of the study is 83.2 months. The histograms regarding the time till the event for the different cause of interests are depicted in Figure 3.2 and 3.3.

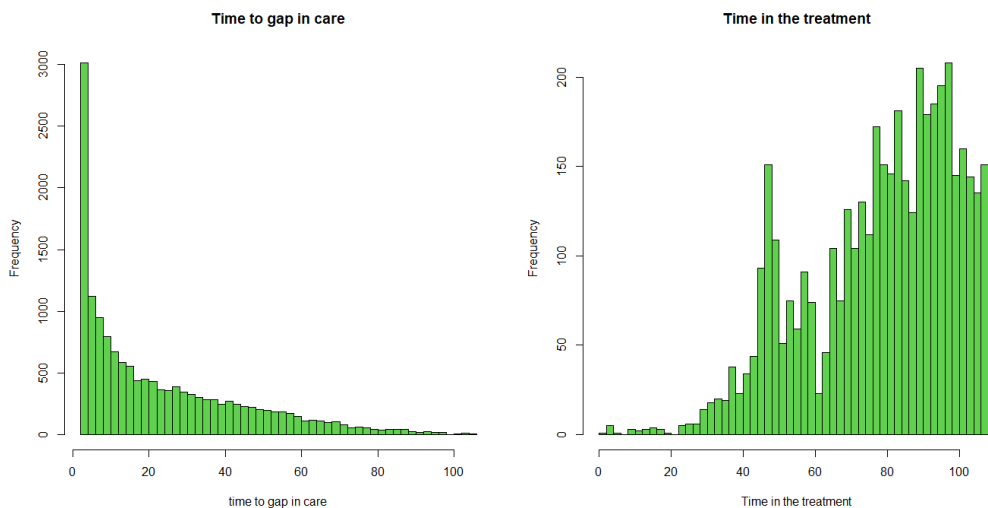


Figure 3.2: The corresponding time-spans until gap in care (right diagram) and until the end of treatment (left diagram).

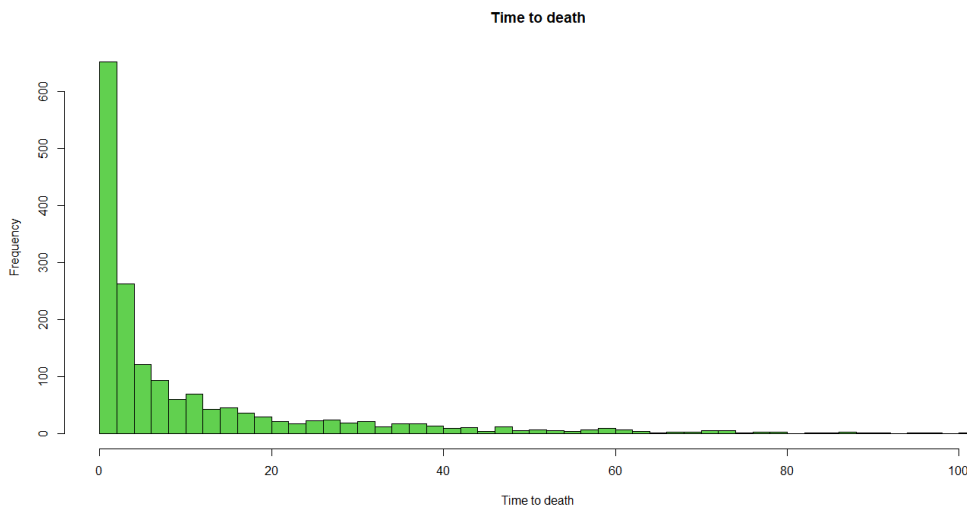


Figure 3.3: The corresponding time-spans until a patient to die.



Via inspection of the aforementioned histograms (Figure 3.2 and 3.3), it is easy to be observed that a lot of patients are either deceased or have gap in care in the first months of the program. The reason is that the treatment in the first months has not yet managed to change the course of the disease and either patients have lost faith or they have passed away because the effects of the treatment were not yet apparent. Month by month the percentage of deaths and people that have a gap in care progressively reduces thanks to the treatment. It is a clear sign of the importance of this kind of programs in the battle against the virus and highlights the need for a larger assistance in the battle against the disease.

Very important, as it has been already mentioned above, for the understanding of the dataset, are the levels with which the health authorities have labeled the patients. In the raw dataset the levels are five and for an easier understanding, the levels one and two have been merged under the naming “medium stage” and the levels three, four and five under the naming “crucial stage”. From the total number of patients, 58% of them are in a medium stage of the disease while 42% of them are in a crucial stage.

Another significant variable is the one that counts the CD4 cells at the start of the antiretroviral treatment. HIV destroys the CD4 cells and as a result the immune system of the patient is not capable of coping with infections. In general, if CD4 cells count is above 500 then everything is considered to be normal. On the other side if it is below 350, then it is a sign that HIV has damaged the immune system of the patient. In the histogram in Figure 3.4, it is obvious that the majority of the patients have CD4 cells lower than the normal and in more than 6000 patients, the count is less than 100, showing that the risk of serious infections becomes much higher and the contribution of the health authorities even more important.



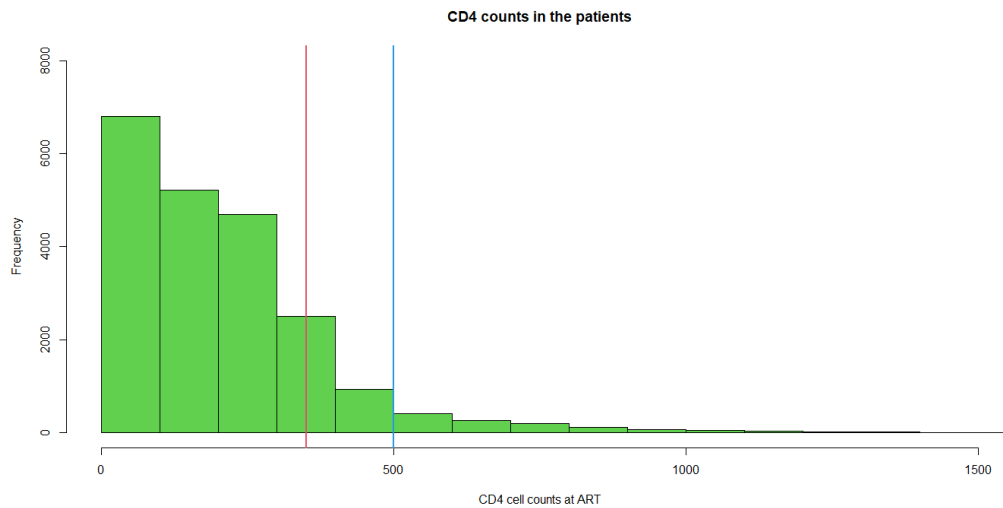


Figure 3.4: The CD4 distribution of the subjects that are in the dataset, depicting that the majority of the patients have less than 500 CD4 cells.

All the aforementioned descriptive statistics for the variables that will be used in the next models are being concentrated in Table 3.4

Cause of failure				
	In Care(N=4473) n(%)	Disengagement (N=4430) n(%)	Death (N=1731) n(%)	Missing (N=10673) n(%)
Gender				
Female	2843 (63.5)	2823 (63.7)	864 (49.9)	7093 (66.4)
Male	1630 (36.5)	1607 (36.3)	867 (50.1)	3580 (33.6)
HIV Status disclosed				
Yes	1787 (40)	1625 (36.6)	485 (28)	4020 (37.6)
No	2686 (60)	2805 (63.4)	1246 (72)	6653 (62.4)
Place of treatment				
Health Center	1137 (25.4)	1000 (22.2)	433 (25)	2065 (19.3)
District Hospital	2250 (50.3)	2362 (53.7)	927(53.6)	6691 (62.6)
Regional Hospital	1086 (24.3)	1068 (24.1)	371(21.4)	1917 (18.1)
	Median (IQR)	Median (IQR)	Median (IQR)	Median (IQR)
Age	39.11 (32.7,46.9)	35.04 (28.4,42.2)	38.83 (32.1,47.9)	35.41 (28.8,43.3)
CD4	199 (101,295)	173 (73,284)	66 (21,167)	183(81,290)

Table 3.4: Table that depicts descriptive statistics for significant variables of our dataset. It splits into categories, depending on the cause of failure and whether is missing or not.

3.3 Pairwise comparisons between the variables

An important “ingredient” in order to understand the problem is to examine the pairwise comparisons between some of the variables. As a result, a small analysis will take place, looking at specific significant relationships between variables. The approach will have to do with the comparison between the patients that have gap in care and the rest of the dataset. The interest has been given to this specific new variable (i.e., dropouts) because of the fact that the variable of the dropouts includes all of the disengagers that is one of the main issues in this analysis.

In the received dataset, as it has already been mentioned, the variable “*orstatus*” provides the status of the people that have gap in care and the health authorities have conducted a tracing procedure. As it has been already mentioned, it was observed that from the 6089 patients that have been traced through outreach, 1659 of them were already dead. Moreover, it should be again highlighted that in these 6089 patients, double sampling will be carried out, since this subgroup is a subset of the sample who contains all the patients who have been considered as disengagers and of course deaths in this subgroup are already known. As a result, they can be representative for the rest of the dropouts and their estimation can be closer to the real value. In the creations of the models in the following pages, the cause of failure for the patients that have not been outreached will be considered as a missing value in order to be estimated.

One of the main interests is to check whether there are differences between the patients that have dropped out and the patients that have not. In general, all the patients that have still a gap in care are considered as dropouts plus the patients that were considered as disengagements but after outreach, it was proved that they had lost their lives. As a result, the dropout category also includes unreported deaths. Testing the association between the continuous variable of age and the categorical variable for the patients that have been disengaged from treatment, the result is that there is evidence of a difference between the groups. It seems that the group of patients that have not been dropped out is older than the other group (the median of the dropout group is 35.57 and the non-dropout group 39.11 with $p\text{-value} < 0.05$). The table that will be presented in the next pages (please refer to Table 3.5), as well as the box-plot in the appendix that depicts the difference between the two medians, highlights the aforementioned description that the age plays a significant role in the occurrence of a dropout.

The variable that counts the CD4 cells at the start of the treatment is the other

important quantitative variable that need to be examined regarding the dropouts. Either via inspection of the box-plot (Please refer to the Appendix) or with the assistance of the Wilcoxon test, since no normality is satisfied and the variance is not equal, it is clear that the difference between the two groups is statistically significant. Patients that have more CD4 cells are more likely to continue the procedure and stay in care. It is logical since, as we have mentioned it before, patients with lower amount of CD4 cells are in a more critical situation, namely weaker and as a result they may skip the treatment because they believe that it does not worth staying in the treatment. The group with the non-dropouts have a higher median CD4 count at enrollment (199 versus 168 in the other group; p -value <0.01). Another very important variable in the understanding of the problem is time until an outcome of interest to be expressed. The mean for the group of the patients that had a gap in care is much smaller than the respective group of the patients that had not (13.95 the former and 82.82 the latter). There is statistical significance between the two groups in terms of time with the corresponding p -value to be smaller than 0.01.

Moreover, it is interesting to see if the two genders have the same tend to dropout. Using the Fisher's exact test, because of the fact that the two variables in scope are categorical, the conclusion is that we do not reject the null hypothesis that the dropouts are independent of the gender, since the p -value=0.519. The same results can be given using other tests like the Pearson's χ^2 independence test. This statistical significance is telling us that the gender does not influences the possibility of dropout and being traced from the patient and specifically the odds of being a dropout patient and female is 1.02 times that for a male person, namely a man is less possible to dropout in general. Furthermore, one indication that there are not significant differences is that the confidence interval does not contain one (0.95– 1.09).

Another important association is the one between the dropout variable and the variable that depicts the kind of the center where the care took place. The initial interest will check whether the health center plays a role in the probability that a patient will become a disengagement. For the variable "level", two groups will be created. The one for the patients with their care in a Health center and the other for the other possible choices (district hospital or university hospital). It seems that there is statistical significance between the two groups since p -value <0.05 . In accordance with the Fisher's Exact test for count data, the possibility for a patient to dropout and the treatment center to be a Health center is 0.76 times the possibility to dropout in other centers.

Our main interest so far was to highlight the differences, so we focused on the



comparisons that will be important for our understanding and later for the creation of the model. All the comparisons that were mentioned were statistically significant, except the gender variable. Please refer to Table 3.5 for a comprehensive view of the results.

Characteristic	Patient Subgroup			p-value	Odds ratio
	Total	Dropouts	Not Dropouts		
	N=21307	N=16773	N=4534		
Gender					
Female	13623	10743	2880	0.519	1.02
Male	7684	6030	1654	0.519	1.02
Center of treatment					
Health center	4635	3481	1154	<0.01	0.76
Other	16672	13292	3380	<0.01	0.76
CD4 count (cells/ml)					
Median	36.38	35.57	39.11	<0.01	-
Age					
Median	176	168	198	<0.01	-
Time					
Median	22.92	13.95	82.85	<0.01	-

Table 3.5: Patient characteristics and comparison between the outreached and non-outreached patients.

As it has been already mentioned before, the cause of failure (i.e., death or disengagement) is missing for 10673 patients that have not been outreached and they are considered as disengagements. For the sub-sample of the people that

have been outreached and have expressed the outcome of death or the outcome of disengagement, a Cox proportional hazards model $\pi(W, \kappa_0)$ could be constructed in order the probability of a patient to express one of the outcomes of interest to be calculated. That approach later will help us to estimate the probability of every cause of failure in the non-outreached patients.

Three different approaches will be followed in this chapter in order to compare the results, before implementing the Bayesian procedure in the next chapter. The main interest is on the patients that have gap in care and have not been outreached since they are considered as missing values and as a result the analysis and the following models will focus on them. However, the models will be constructed from the whole dataset of the 21307 patients and as a result, they will give a general insight regarding the two important aspects of the analysis: the hazard ratios and the significance of the covariates in the analysis.

Last but not least, it should be noted that a transformation of the variables of age and CD4 was implemented in order more meaningful results to be given. Instead of calculating whether the difference of one cell count is significant, something that will not give important information regarding the variable, the variable will be cut into categories. In the next chapters, we will examine whether the change of 100 CD4 cells plays important role in the corresponding models. Respectively, the age variable will be split into groups where each group includes patients that are in the same decade. The variable time will still use months as point of reference.

3.4 Complete case analysis with Cox proportional hazards model

The first way that will be introduced is a Cox proportional hazards model since it can estimate better with the presence of time in comparison, for example, from a simple logistic model. In this first model complete case analysis will take place, namely all patients with missing cause of failure will be omitted. The rest 10634 patients, whose outcome has been outreached (either disengagement or loss of life) or they are considered as right censorings, are included in the analysis. If our



analysis only included the patients that have expressed one of the two competing risks and not the ones that have not expressed an outcome during the treatment (right censorings), then the whole procedure will be considered biased, since we do not include patients with important information (complete data). Since the missing at random assumption holds, the formula 2.4 tells us that the results from the complete case analysis can be representative for the subset of patient whose outcome is not known.

Furthermore, the covariates that our model will have, are the following, namely gender, age, location where the treatment is held, CD4 cell count at the time of enrollment, time from the start of treatment and whether or not HIV status is disclosed. Other covariates such as dates of the visit in the health center, identification number of the center and the patient and if the patient is transferred in another center have been omitted due to the fact that they contain a lot of missing values and the estimation will be less precise. Due to the fact that the Cox proportional hazards model estimates the hazard ratios of a cause of failure, two models will be fitted, each for every cause of failure, and a new variable will be introduced. This variable will be considered as a binary status variable. If the interest lies upon loss of life, then the variable will take the value one and zero otherwise, while if the interest lies upon the outcome of disengagement, then it will take the value one and zero otherwise. In the formula 3.1 below, the reference level for the gender will be the female patients, for the variable regarding HIV status the reference level will be the no answer. Regarding the variable about the place of the treatment, since there are three categories, the reference level will be the Health center in both variables, where in level1 the comparison between Health center and district hospital is taking place and in level2 the comparison between Health center and the university hospital. The formula that will be used for the Cox proportional hazards model is the following:

$$h(T|X) = h_0(T) * \exp(\beta_1 * I(\text{sex} = \text{male}) + \beta_2 * I(\text{hivstatus} = \text{yes}) + \beta_3 * \text{Age} + \beta_4 * \text{CD4} + \beta_5 * \text{Level1} + \beta_6 * \text{Level2}), \quad (3.1)$$

,where T is the time and $h_0(T)$ the baseline hazard function, which represents the hazard when all covariates are zero (reference level).

After fitting the model, it has been observed that the three overall tests (likelihood, Wald, score) have given p-values smaller than 0.001 indicating that there is strong evidence that the model is superior that the null model without covariates. In this analysis, these tests assess the validity of the null hypothesis, namely



whether the betas are zero. The results of the two Cox proportional hazard models are portrayed in the Tables 3.6 and 3.7.

Covariates	Hazard ratio	Confidence interval	Standard error	P-value
Age at enrollment (per 10 years)	0.750	(0.728,0.773)	0.015	<0.001
Gender (Female vs Male)	1.071	(1.006,1.141)	0.032	0.030
CD4 Cell count (per 100 cells/ μ l)	0.984	(0.965,1.003)	0.009	0.099
Place of the program (1)	1.084	(1.006,1.167)	0.037	0.032
Place of the program (2)	1.052	(0.965,1.147)	0.044	0.246
HIV status disclosed (no vs yes)	1.004	(0.944,1.068)	0.031	0.883

Table 3.6: Estimates of the covariates and standard errors for the first model (Cox proportional hazards model) when disengagement is the outcome of interest.

Covariates	Hazard ratio	Confidence interval	Standard error	P-value
Age at enrollment (per 10 years)	1.077	(1.030,1.126)	0.022	0.001
Gender (Female vs Male)	1.364	(1.239,1.503)	0.049	<0.001
CD4 Cell count (per 100 cells/ μ l)	0.611	(0.582,0.641)	0.024	<0.001
Place of the program (1)	1.028	(0.917,1.153)	0.058	0.630
Place of the program (2)	0.795	(0.692,0.915)	0.071	0.001
HIV status disclosed (no vs yes)	0.679	(0.611,0.755)	0.053	<0.001

Table 3.7: Estimates of the covariates and standard errors for the first model (Cox proportional hazards model) when death is the outcome of interest.

Upon inspection of Tables 3.6 and 3.7, if the outcome of interest is the disengagements, the variables age and gender are statistically significant along with the one level regarding the place of the treatment (health center and district hospital). On the other side, if the outcome of interest is loss of life, then all the variables in the model are statistically significant except of the first level of the place of the treatment. For the interpretation let's take the variables HIV status disclosed, gender and age as examples.

Starting from the variable "HIV status disclosure", the hazard ratio for the competing event of death is 0.679 and this is an indication that there is a strong relationship between this variable and decreased risk of death (the predictor is protective). In other words, if someone, and supposing that all the other covariates are constant, has stated yes, namely she/he has stated that has HIV, then the hazard of death is reduced by 32% or, in other words, the hazard reduces by a factor of 0.678. The confidence interval for that variable lies between 0.611 and 0.755, so the equality between the two groups regarding that variable cannot be in place. On the other side, if disengagement is the outcome of interest, then the hazard ratio is 1.004, indicating that a patient that has stated yes has lower probability of not

been disengaged. Although the hazard ratio lies near one and the received p-value is bigger than 0.05 (p-value=0.883), so it seems that there is no evidence for an association between the hazard of disengagement from care and whether she/he answered yes or not in the question regarding HIV status disclosure.

Looking at the gender variable, it can be observed that a male has 36% greater hazard of dying than a woman, presupposing that all the other variables are constant, while the corresponding hazard ratio for disengagement indicates that female patients are more likely not to express this outcome of interest than men. Likewise, we see that the hazard ratio of death in the variable of age is 1.077, with a confidence interval (1.030,1.126), while for disengagement is 0.750, where the confidence interval is narrow (0.728,0.773). The confidence interval does not include one and as a result there is an indication that contribution of that specific covariate to the model is statistically significant. Regarding the outcome of death, holding the other covariates constant, with an increase of age group the risk of death increases by 7.7%. This is logical. Since as the patient gets older, its immune system is not so strong. On the other side, it seems that if the difference between two patients is only one age group, then it has 25% less risk of have a gap in care. Also, in this situation, the immaturity of younger people can play a significant role in their decision to leave the treatment, even if they know that this virus can kill them. Of course, it should be mentioned again, that in this model is created with the complete cases and as a result, there are discrepancies from the actual results.

For the aforementioned models, a test for non-proportional hazards has been conducted using the function `cox.zph` in R and the test is statistically significant for all the variables except of the place of the program and HIV status when the outcome is loss of life and except of the place and the age when the outcome is disengagement from treatment. As a result, the assumption of proportional hazards for the model has been violated, result that can be also proved by the global test for non-proportional hazards, which is statistically significant. The same violation is met even if the variables are stratified. However, the stratification approach will not be followed because of the fact that the hazard ratio for the stratified variables cannot be obtained due to the fact that the baseline hazard has absorbed that effect. Moreover, it should be noted that the aforementioned violation is supported by plots of the Schoenfeld residuals. After the graphical inspection of the Schoenfeld residuals regarding the outcome of death (please refer to Appendix, where the Schoenfeld residuals for age at enrollment and CD4 count are portrayed), it is clear that a pattern with time is in place, since as time passes by there is a change of the graph.



3.5 Multiple Imputation and Cox model using the predicted probabilities of logistic regression

The next approach will be a more complicated one. The first step in this approach will be to fit a logistic model for all the patients that have been outreached from the health authorities. The right censorings will not be included because it is known that the patients whose outcome is missing are not right censorings. As it has already been referred, all the patients that have not been outreached are wrongly acclaimed as disengagers. Given the fit of the logistic model, where the patients that are included in the Expensive gold standard event type ascertainment will be used, the predictive probabilities for patients with missing cause of failure will be received and will be used for the imputation of these missing outcomes. This probability will be the probability of success on each trial in the Bernoulli distribution that will be utilized in order the missing values to be simulated. If, for example, a patient that has not been outreached has a predicted probability of 0.6, then this probability will be used for the imputation of the missing value, thanks to the Bernoulli distribution, and the later survival analysis.

In order to acquire a better and more precise result, 100 different data sets will be created and used. Due to the fact that these results have been extracted from a simulation, the number of deceased people cannot be considered as exact. Moreover, due to the fact that simulation has taken place, there may be small differences if the procedure will be followed again. The next step that will take place, is the fit of 100 Cox proportional hazards models, as the number of the simulated data sets. The same variables will be used as in the previous approaches since there was no actual improvement with the scale of the variables or the introduction of logarithms and square roots.

With the fit of 100 different models, 100 different values for every covariate of the two models (one for death as outcome and one for disengagements) were estimated. In order to have a general and precise estimation of every covariate of interest for the competing risks, the mean and the standard deviation of these variables were estimated. This approach is considered better from the aforementioned ones, since the missing values are simulated from the predicted values thanks to the binomial distribution. This aforementioned procedure will be repeated using the bootstrapping approach in order to get more precise results regarding the standard errors. In general, bootstrapping is a procedure where a dataset is re-sampled



with replacement and creates many simulated data sets. Following this procedure, standard errors, confidence intervals and hypothesis testing can be estimated more precise than the previous approaches. In general, the Bootstrap approach assumes that the original dataset that will be used is representative of the population and as a result non-parametric estimations are possible, namely it does not need specific distributional assumptions to be satisfied.

In our example, for the estimation of the standard errors, 100 different simulated data sets will be sampled with replacement from the initial dataset and then the aforementioned procedure with the creation of 100 different data sets, the estimation of the cause of failure using the predicted probability and the fit of the Cox proportional hazards models will take place for every one sampled dataset. From these bootstrapped data sets, a vector of 100 values for every covariate of interest for the two competing risks will be created, where every value has been estimated from the average of the values of the covariates from the 100 data sets. It should be mentioned that the standard deviation of the 100 multiple imputation estimates is the standard error estimate. Using this approach, more precise standard errors and confidence intervals can be calculated. Lastly, it should be noted that the assumption of proportional hazards is also being violated with this approach (for example for the gender variable).

To sum up, the steps that were applied in this approach are the following:

- 1) Isolate the sub-sample that contain the outreached patients and apply a multiple logistic regression.
- 2) Obtain the predicted probabilities for the not outreached patients and simulate 100 times the outcome of interest based on the binomial distribution, using as probability the obtained predictive probability from the logistic regression.
- 3) Having created 100 different outcomes, merge the sub-samples of outreached, not outreached and right censorings and fit 100 different Cox proportional hazard models for deaths and 100 for disengagements in order later to obtain the hazard ratios and the p-values.
- 4) Exponentiate the results of each variable in the two models in order to get the hazard ratios and take the mean value from them.
- 5) Since the results have been received, use the bootstrap approach. Sample 100 different data sets with replacement from the initial dataset, that includes both missing values and precise results regarding the cause of failure and follow the steps 1-4 (create a function in order to be computationally easier).
- 6) From a vector that contains the 100 values for the same variable (each of the values is the mean from step 4), calculate the standard error as well as the 95% confidence intervals and the p-values.



It should be noted that in the multiple logistic regression that is used for the estimation of the predicted probabilities, a model with piecewise linear effect of time with a change in slope at 12 months is used. Upon inspection of Figure 3.2 (time until disengagement) and Figure 3.3 (time until death), it is clear that after the first year of the treatment, there is a steep fall of the patients with appearance of one of the two possible outcomes of interest. As it was referred, this is logical since as time passes by, the effects of the treatment will be apparent. In order the model to be fitted, it is necessary a dummy variable to be created in order the times lower than 12 months and the values higher to be differentiated.

Furthermore, regarding p-values, they can be estimated from the Wald test, namely the formula $W = \frac{\beta}{S.E}$, where β is the point estimate of each covariate of interest and the S.E is the standard error that has been granted after bootstrap for every covariate of interest. P-values are equal to the formula $2 * [1 - \Phi(|W|)]$, where $\Phi(|W|)$ is the cumulative distribution function (CDF) of the standard normal distribution.

Please refer to Tables 3.8 and 3.9 for the results of the approach after multiple imputation and bootstrapping approach and the comparison with the Cox proportional hazards model that was portrayed in the previous chapter.



Covariates	Hazard ratio	Confidence interval	Standard error	P-value
Age at enrollment (per 10 years)	0.795	(0.775,0.814)	0.010	<0.001
Gender (Female vs Male)	1.043	(1.002,1.083)	0.020	0.035
CD4 Cell count (per 100 cells/ μ l)	1.047	(1.035,1.058)	0.005	<0.001
Place of the program (1)	1.145	(1.091,1.198)	0.027	<0.001
Place of the program (2)	1.035	(0.967,1.102)	0.034	0.316
HIV status disclosed (no vs yes)	1.133	(1.092,1.173)	0.020	<0.001

Table 3.8: Estimates of the covariates and standard errors for the Multiple Imputation model (Cox proportional hazards models estimations for disengagements).

Covariates	Hazard ratio	Confidence interval	Standard error	P-value
Age at enrollment (per 10 years)	1.088	(1.047,1.128)	0.020	<0.001
Gender (Female vs Male)	1.338	(1.260,1.415)	0.039	<0.001
CD4 Cell count (per 100 cells/ μ l)	0.681	(0.632,0.716)	0.021	<0.001
Place of the program (1)	1.069	(0.969,1.168)	0.051	0.190
Place of the program (2)	0.766	(0.632,0.899)	0.068	<0.001
HIV status disclosed (no vs yes)	0.749	(0.660,0.837)	0.045	<0.001

Table 3.9: Estimates of the covariates and standard errors for the Multiple Imputation model (Cox proportional hazards models estimations for deaths).

Upon inspection of Tables 3.8 and 3.9, it seems that there are differences from the previous approach and this can be explained by the fact that imputation was applied in this procedure. The majority of the variables in both models have kept the same factor (protective or not) except of the count of CD4 cells when disengagement is the outcome of interest. Following this approach, the increase of CD4 cells seems to be a hazard factor in the deterrence of the outcome of interest unlike the complete case model, where it acts as a protective factor. Lastly, there are differences in the p-values, however this can be explained due to the large number of missing values that were imputed.

3.6 MPPL approach

The final frequentist approach in this chapter follows the *Maximum pseudo-partial-likelihood estimation (MPPLE)*, the main frequentist way in this thesis to solve this problem. This approach fits semi-parametric proportional hazards model for competing risks data with missing cause of failure. Similarly to the multiple imputation in the previous chapter, it follows the same idea of imputing the missing cause of failures. Even though the model is considered as semi-parametric, the probability for the cause of failure is calculated thanks to a parametric model (multiple logistic regression). As it has already been referred, this approach gives more statistical and computational efficiency. The steps that need to be done are the following:

- 1) Create a variable that defines whether or not the cause of failure is known.
- 2) Fit a logistic model in the complete cases (use the outreached patients) and estimate the predicted probabilities for each cause of failure of the initial dataset. The same approach with the Multiple Imputation procedure regarding the split of the time variable is used.
- 3) Estimate the weights that are calculated with formula 2.9 from the second chapter.
- 4) For the observations where the event type is missing, remove weights from the risk sets, namely patients with cause of failure.
- 5) Fit two Cox proportional hazard models, namely one for each cause of failure, using the weights that were created in step 3.
- 6) Estimate bootstrap standard errors or implement the closed-form standard error estimator proposed in *Bakoyannis et al. (2020)*.
- 7) Plot the Cumulative Incidence Function.

Moreover, this approach can be followed using the *ClusteredMMPLE* package and the function *ccr_smreg*. This function provides the confidence bands for the residual process for goodness of fit evaluation. Using 1000 multiplier replications for computing the latter bands is currently computationally intensive and requires more efficient computation. In this thesis, it was used in order to visualize the cumulative residual process through time (please refer to Figure 3.5). Please refer to Tables 3.10 and 3.11, where the estimates of that approach for the two outcomes of interest are portrayed.



Covariates	Hazard ratio	Confidence interval	Standard error	P-value
Age at enrollment (per 10 years)	0.798	(0.784,0.811)	0.007	<0.001
Gender (Female vs Male)	1.037	(1.007,1.066)	0.015	0.009
CD4 Cell count (per 100 cells/ μ l)	1.039	(1.029,1.048)	0.005	<0.001
Place of the program (1)	1.164	(1.136,1.191)	0.014	<0.001
Place of the program (2)	1.045	(1.019,1.070)	0.013	0.002
HIV status disclosed (no vs yes)	1.144	(1.122,1.165)	0.011	<0.001

Table 3.10: Estimates of the covariates and standard errors for the third model (Marginal Regression-Disengagements).

Covariates	Hazard ratio	Confidence interval	Standard error	P-value
Age at enrollment (per 10 years)	1.102	(1.082,1.121)	0.010	<0.001
Gender (Female vs Male)	1.339	(1.311,1.366)	0.014	<0.001
CD4 Cell count (per 100 cells/ μ l)	0.676	(0.650,0.701)	0.013	<0.001
Place of the program (1)	1.040	(0.998,1.081)	0.021	0.061
Place of the program (2)	0.791	(0.722,0.859)	0.035	<0.001
HIV status disclosed (no vs yes)	0.781	(0.743,0.818)	0.019	<0.001

Table 3.11: Estimates of the covariates and standard errors for the third model (Marginal Regression-Deaths).

In the case of death, only the variable that compares a health center and a district hospital can be considered as not statistically significant for some levels of significance. In case of disengagements, there is evidence that all the variables are important in the model. Furthermore, the results seem to be logical. For example, with the increase of CD4 cell counts, the hazard for a patient to lose her/his life plunges, since the hazard ratios is 0.67. Moreover, younger patients seem to have smaller hazard to die in comparison to the older ones but if disengagement is the outcome of interest, then they have higher hazard ratios than the older patients. The latter result can be explained due to the immaturity of younger people that do not take a lot of situations as serious as the older ones. Another example, is the gender. If death is the case, then it is clear that men are in higher danger than women with p-value for that variable lower than 0.01. That could be explained from the generic though that the immune systems of women are more resistant. Last but not least, also in the MPPL approach, the CD4 cell count variable is a

risk factor contrary to the complete case analysis.

The cumulative residual process for the evaluation of the aforementioned parametric logistic model with the 95% goodness of fit (the grey area) does not provide evidence for a lack of fit, even if in the early time-points there is a higher variance of the residuals. These variances however remain always within the 95% confidence band under the null hypothesis so as a result the underestimation of death in the beginning is not statistically significant ($p\text{-value}=1$). After the first year of the program, the discrepancies are smaller and the cumulative residual process has values near to zero. Using this approach, there is no indication of model misspecification and under or overestimation of death. The following plot was obtained thanks to the *ccr_smreg* function.

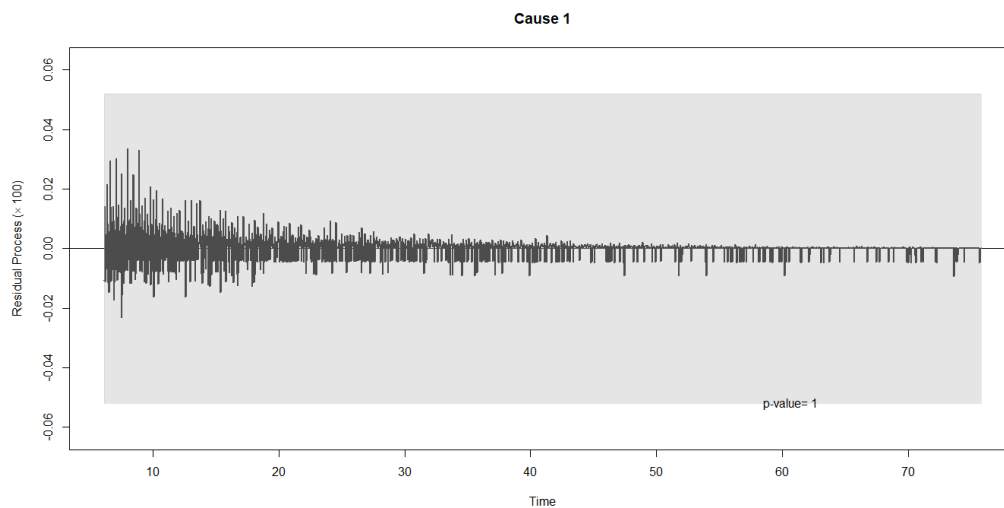


Figure 3.5: Plot of cumulative residual process through time.

Moreover, plots for the cumulative incident functions of each cause through time were calculated.

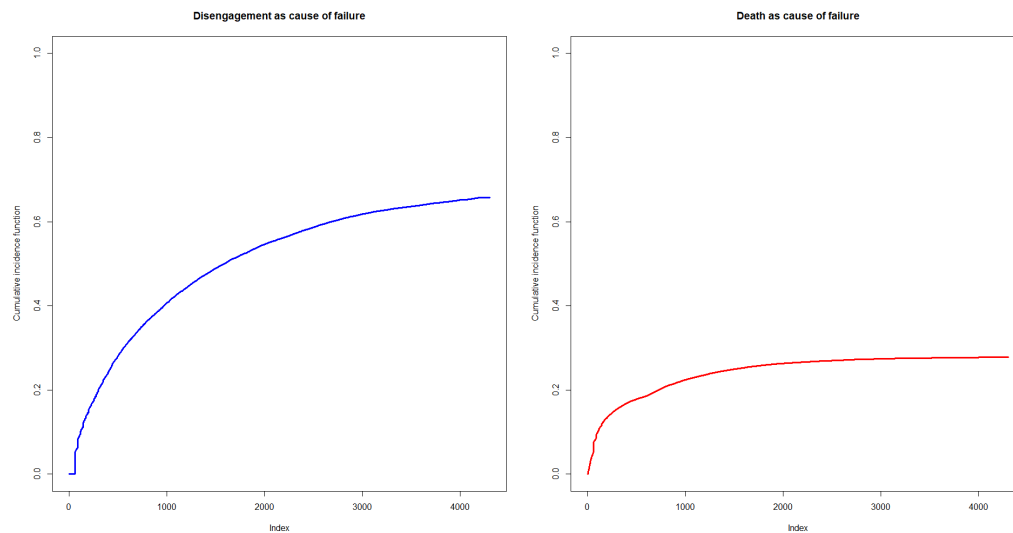


Figure 3.6: Plot of cumulative incidence function through time.

Upon inspection of Figure 3.6 it can be clearly observed that the probability for a patient to die is lower than the probability of a gap in care to occur and lies between 0.2 and 0.3. Furthermore, the probability of death is not significantly increasing in contrast to the corresponding probability of a patient to have a gap in care where the increase is continuous as time passing by. Regarding the disengagement as an outcome, the probability progressively increases and picks approximately at 0.6. It is obvious that the cumulative incidence of death can be considered stable after the first year of the treatment and this is a clear sign that the procedure is effective. On the other side, the continuous increase of the probability of disengagement through time can be explained by various reasons that have not necessarily have to do only with the effectiveness of the treatment.

3.7 Comparison of the models and conclusion of third chapter

In this chapter, the main focus was to understand and evaluate the frequentists approaches for competing risks analysis with missing causes of failure. Very significant is to obtain as small standard error as possible. The Multiple Imputation as well as the MPPL approach have given better results than the first and simple complete case analysis and as a result the former procedures are preferred. In the Tables 3.12 and 3.13 below, a comparison of the hazard ratios and their standard errors of all the aforementioned approaches is portrayed.

Covariates	Hazard ratio (Relative efficiency) complete case analysis	Hazard ratio (Relative efficiency) MI	Hazard ratio (Relative efficiency) MPPL approach
Age at enrollment (per 10 years)	0.750 (4.591)	0.795 (2.040)	0.798 (1)
Gender (Female vs Male)	1.071 (4.551)	1.043 (1.777)	1.037 (1)
CD4 Cell count (per 100 cells/ μ l)	0.984 (2.250)	1.047 (1.012)	1.039 (1)
Place of the program (1)	1.084 (6.984)	1.145 (3.719)	1.114 (1)
Place of the program (2)	1.052 (11.455)	1.035 (6.840)	1.045 (1)
HIV status disclosed (no vs yes)	1.004 (7.942)	1.133 (3.305)	1.144 (1)

Table 3.12: Comparison of all the models from the third chapter (Disengagements).

Covariates	Hazard ratio (Relative efficiency) complete case analysis	Hazard ratio (Relative efficiency) MI	Hazard ratio (Relative efficiency) MP-PLE approach
Age at enrollment (per 10 years)	1.077 (4.840)	1.088 (4)	1.102 (1)
Gender (Female vs Male)	1.364 (12.250)	1.338 (7.760)	1.339 (1)
CD4 Cell count (per 100 cells/ μ l)	0.611 (3.408)	0.681 (2.609)	0.676 (1)
Place of the program (1)	1.028 (7.628)	1.069 (5.897)	1.040 (1)
Place of the program (2)	0.795 (4.115)	0.766 (3.774)	0.791 (1)
HIV status disclosed (no vs yes)	0.679 (7.781)	0.749 (5.609)	0.781 (1)

Table 3.13: Comparison of all the models from the third chapter (Deaths).

Moreover, as it has already been referred, the last approach gives more precise results with lower standard errors on average than the other methods. This can be observed thanks to the relative efficiency between the standard errors of the two comparable models, namely thanks to following formula

$$Relative\ efficiency = \frac{standard\ error_{MI}^2}{standard\ error_{MPPLE}^2}, \quad (3.2)$$

In this aforementioned formula 3.2, the relative efficiency corresponds to the variance of the estimate from a given method versus the variance of the corresponding MP-PLE estimate. Upon inspection of the Tables 3.12 and 3.13 and thanks to the relative efficiency, it is clear that in both outcomes of interest, the last approach is the best one, especially when death is the main interest. The smallest difference in the standard errors between the models is between the disengagers of the model with multiple imputation and the model with MP-PLE. Both imputation methods are more efficient than the complete case analysis. In general, all the analyses, indicate the fact that the estimation of the missing data is

crucial, because of the fact that the missing values are a lot, there are a lot of differences between the multiple imputation, the MPPL methods and the complete case analysis.

Last but not least, since the MPPL approach is a Cox proportional hazards model with the “corrected” causes of failure for the non-outreached to be imputed from the predicted probabilities of a logistic model, it is interesting to see if there is difference in the estimates between this model and a Cox proportional hazards model with the misspecified data. Tables 3.14 and 3.15 indicate that due to the fact that our dataset had a lot of missing values, the hazard ratios between the proposed imputed approaches and the misclassified model differ substantially. All the variables have notable differences, although the biggest ones can be spotted in the variables that describe the gender and whether the HIV status is disclosed or not.



Covariates	Hazard ratios (Cox proportional hazards model)	Hazard ratios (multiple imputed model)	Hazard ratios (MPPLE model)	Hazard ratios (Misspecified model)
Age at enrollment (per 10 years)	0.750	0.795	0.798	0.838
Gender (Female vs Male)	1.071	1.043	1.037	1.072
CD4 Cell count (per 100 cells/ μ l)	0.984	1.047	1.039	1.002
Place of the program (1)	1.084	1.145	1.164	1.175
Place of the program (2)	1.052	1.035	1.045	1.006
HIV status disclosed (no vs yes)	1.004	1.133	1.144	1.095

Table 3.14: Comparison between the proposed models and a model with the misspecified data (disengagements).

Covariates	Hazard ratios (Cox proportional hazards model)	Hazard ratios (multiple imputed model)	Hazard ratios (MPPLE model)	Hazard ratios (Misspecified model)
Age at enrollment (per 10 years)	1.077	1.088	1.102	1.128
Gender (Female vs Male)	1.364	1.338	1.339	1.473
CD4 Cell count (per 100 cells/ μ l)	0.611	0.681	0.676	0.614
Place of the program (1)	1.028	1.069	1.040	0.843
Place of the program (2)	0.795	0.766	0.791	0.814
HIV status disclosed (no vs yes)	0.679	0.749	0.781	0.964

Table 3.15: Comparison between the proposed models and a model with the misspecified data (deaths).

In the next chapter the main focus will be the creation of a model using a suitable parametric Bayesian approach to competing risks data with missing event types. The choice of an appropriate prior distribution can play a huge role in the analysis and an attempt will take place in order to generalize the results of a chosen sample to the whole dataset.

Chapter 4

Bayesian approach for competing risk data with missing event types

4.1 Selection of the prior distribution

In this chapter, the focus will be on the creation of a Bayesian methodology for competing risk data with missing event types. As we saw on the previous chapter, the model that uses the MPPLE approach delivers better results than the other simpler models. Since, one of the intentions is to estimate the cause of failure for the not outreached patients, these missing values will be treated as parameters and priors will be assigned to them. Since the possible causes of failure are two, a binomial distribution can be used as a prior distribution for the response, namely the missing values. Due to the fact that the not outreached patients can either express death or disengagement as an outcome of interest, for the estimation of the outcome multiple logistic regression will be used in one of the approaches. Furthermore, for the estimation of the corresponding parameters, the statistical software program JAGS (Just Another Gibbs Sampler) will be used. JAGS is using Markov chain Monte Carlo (MCMC) methods such as Gibbs sampling algorithm in order to converge to the corresponding estimate. JAGS was preferred instead of other Bayesian modeling packages for its speed and its compatibility.

It should be repeated that the Missing at random (MAR) assumption is fol-



lowed in the dataset. This can be observed by the fact that the missing data have a dependence with other variables of the dataset, as it has been referred above, for example, male patients with lower CD4 cell count and younger age are more likely to be disengaged than others. There are 10673 patients whose outcome is considered as missing and for these values, prior distributions need to be defined as it was referred above. The sample of the disengagements will also be split into two sub-samples, one with the patients that have been outreached and one with these that have not been outreached and are the missing values. The prior belief for the missing outcomes can be “extracted” from the outreached patients or non-informative priors can be used. Although, the prior belief should not be characterized as subjective.

But what is the most suitable approach to determine the prior predictions and the prior distribution? There are common priors that could be used but the focus will not be concentrated into all of them. Low informative priors such as a simple uniform prior or the Jeffrey’s prior could be used. Jeffrey’s prior is analogous to the square root of the determinant of the matrix of Fisher information. On the other side, the uniform prior basically considers that all the parameters in the parameter space have the same probabilities. A uniform $\theta \sim U(0, 1)$ can equivalently be written in terms of a beta distribution as $\theta \sim Beta(1, 1)$. Obviously, another and more suitable choice would be the conjugate prior of the binomial distribution, namely the beta distribution. It is called conjugate because prior and posterior follows the same distribution. In other words, using this approach, the posterior distribution would be known. This prior can give strong prior knowledge to our model and could give proper estimates. Last but not least, a non-informative prior can be considered as a more objective approach and can be preferred in cases where the credibility in the prior belief cannot be considered as high.



4.2 Complete case analysis using the Weibull baseline hazard function

The first approach that will be followed is fitted thanks to the Weibull distribution. As it was mentioned in the first chapter, this model can be considered as a generalization of the exponential model, although it is considered to be more flexible due to the fact that the hazard rate function in the Weibull model is not constant with time ($h(t) = \lambda \alpha t^{\alpha-1}$) as it is in the case of the exponential ($h(t) = \lambda$). A Weibull model has advantages in comparison to the exponential one since it can be expressed either as an accelerated failure time (AFT) model or as a proportional hazard model. The flexibility allows the Weibull model to capture a wider range of survival patterns, including increasing, decreasing and constant hazard rates.

In comparison to the semi-parametric Cox model, there are reports such as *Application of Weibull model for survival of patients with gastric cancer (Zhu et al., 2007)*, that propose that the Weibull model gives more precise results than the Cox proportional hazard model. Moreover, the analysis of a semi-parametric model in the Bayesian framework may cause computational difficulties and that is why it will not be preferred. In general, the complexity of the model structure and the interaction between the parametric and non-parametric components can make the computation of the posterior distribution challenging and time-consuming.

Before starting, there are some “ingredients” in the function Jags that will be used and needs to be referred. These “ingredients” are also important in all the possible packages that have to do with MCMC methods. It is very important to define the `n.thin`, which indicates that the model will keep every `n` value, where `n` is a number that will be defined by the user. All the other values created from the MCMC Gibbs sampler will be omitted. This value is significant, since a good choice of `n.thin` could possibly decrease the autocorrelation between the data and increase the sampling efficiency.

It is very significant to define the number of chains (`n.chains`), which indicates the number of sampled distributions. Using multiple chains serves in the check of MCMC convergence and to obtain more precise estimates of the parameters. Moreover, it is crucial to define the burn in period (`n.burnin`), namely the number of values that will be omitted before starting estimating the posterior distribution. If burn in period is not defined, then all the early values that can be given due



to poor priors can cause bias to the posterior distribution. Last but not least, the number of iterations (n.iter) determines the number of values that are retained for the posterior distribution after burn-in values have been thinned and discarded. For the Bayesian analysis of the model with Weibull baseline hazard function, the thin was set to ten, the burn-in period was set to 1000 values that were omitted, while the number of iterations was set to 10000 and three chains were sampled for that example.

In general, the Weibull model can be defined as

$$Y_i \sim Weibull(\lambda, \alpha), \quad (4.1)$$

where Y_i is the time until the event to be expressed for the patient i , λ is the scale and α the shape of the Weibull distribution respectively. The corresponding cause-specific hazard functions are modelled thanks to the following expression

$$h_k(t|h_{0k}, \beta_k) = h_{0k}(t) * \exp(\beta_1 * I(\text{sex} = \text{male}) + \beta_2 * I(\text{hivstatus} = \text{yes}) + \beta_3 * \text{Age} + \beta_4 * \text{CD4} + \beta_5 * \text{Level1} + \beta_6 * \text{Level2}), \quad (4.2)$$

where the specified Weibull baseline hazard function $h_{0k}(t)$ is the following

$$h_{0k}(t) = \lambda_k \alpha_k t^{\alpha_k - 1}, \quad (4.3)$$

where α_k and λ_k are the shape and scale parameters respectively, while $k=1,2$ are the two possible causes of failure. Moreover, it should be noted that the likelihood of the Weibull model is split into two parts and it depends on whether the given observation is censored or not.

Assuming that the censoring is independent, if the observations are censored then the likelihood function for the Weibull survival model will be the following:

$$L_{\text{censored}}(\beta, \lambda, \alpha) = \prod_{i=1}^n [S(t_i; \beta, \lambda, \alpha)]^{1-\delta_i} = \prod_{i=1}^n \exp(-\lambda t_i^\alpha)^{1-\delta_i} \quad (4.4)$$

If the observations are uncensored then the likelihood function is the following

$$L_{\text{uncensored}}(\beta, \lambda, \alpha) = \prod_{i=1}^n [f(t_i; \beta, \lambda, \alpha)]^{\delta_i} = \prod_{i=1}^n \lambda \gamma_i^{\gamma-1} \exp(-\lambda t_i^{\gamma})^{\delta_i} \quad (4.5)$$

The combined Weibull likelihood is the following

$$L(\beta, \lambda, \alpha) = \prod_{i=1}^n [S(t_i; \beta, \lambda, \alpha)]^{1-\delta_i} \cdot \prod_{i=1}^n [f(t_i; \beta, \lambda, \alpha)]^{\delta_i} = \prod_{i=1}^n \exp(-\lambda t_i^{\gamma})^{1-\delta_i} (\lambda \gamma_i^{\gamma-1} \exp(-\lambda t_i^{\gamma}))^{\delta_i} \quad (4.6)$$

,where δ is the censoring indicator (0 for censored and 1 for uncensored), $S(t_i; \beta, \lambda, \alpha)$ is the survival function and $f(t_i; \beta, \lambda, \alpha)$ is the probability density function.

Non-informative prior distributions will also be used in this model. The betas will follow a $N(0,0.01)$ while the shape and scale parameters will follow a $\text{Gamma}(0.01,0.01)$ and a $U(0,10)$ respectively. Moreover, in order for the results to be granted with lower standard errors and be more precise, the “zero trick” approach was used. As it has already been mentioned, one of the main differences between the Bayesian and frequentist frameworks is that in Bayesian inference, parameters are treated as random variables with their own probability distributions. Prior beliefs about the parameters are specified and these beliefs are updated based on observed data to obtain posterior distributions, while in frequentist inference, parameters are fixed and unknown, but not assigned probability distributions. Point estimates, such as maximum likelihood estimates, are used to estimate the true values of parameters.

We will follow the same approach as in chapter three and we will create two Bayesian models in order to make the comparison between them and compare the Bayesian and frequentist frameworks. The first will be with the complete data, namely we will omit the patients with missing cause of failure, while in the second approach the missing cause of failure will be estimated thanks to the



Bayesian multiple logistic regression model and the results will be used for the estimation of the hazard ratios in the survival model with the Weibull baseline hazard function. In this aforementioned Bayesian logistic model, the outcome of interest for the not-outreached patients will be estimated. For the complete case model, non-informative priors will be used. For the covariances of the variables, normal distribution will also be used with $N(0,0.01)$. The prior probability for the estimation of the outcome (p_0) follows the beta distribution, namely $Beta(1,1)$. Finally, the covariance of the intercept of the logistic model will be the log of odds of the probability (p_0).

Regarding the complete case analysis for the Weibull model, the dataset will include the patients who have either a known cause of failure or are right censorings. For the complete case analysis, please refer to Tables 4.1 and 4.2 in the next page.



Covariates	Hazard ratios	Standard error	Hazard ratios of the frequentist complete case analysis	Standard error of the frequentist complete case analysis
Age at enrollment (per 10 years)	0.741	0.016	0.750	0.015
Gender (Female vs Male)	1.066	0.032	1.071	0.032
CD4 Cell count (per 100 cells/ μ l)	0.991	0.010	0.984	0.009
Place of the program (1)	1.078	0.039	1.084	0.037
Place of the program (2)	1.050	0.043	1.052	0.044
HIV status disclosed (no vs yes)	1.019	0.031	1.004	0.031

Table 4.1: Estimates of the covariates and standard errors for the Bayesian approach for the disengagements (Complete case analysis) and comparison to the frequentist one.

Covariates	Hazard ratios	Standard error	Hazard ratios of the frequentist complete case analysis	Standard error of the frequentist complete case analysis
Age at enrollment (per 10 years)	1.052	0.023	1.077	0.022
Gender (Female vs Male)	1.377	0.050	1.364	0.049
CD4 Cell count (per 100 cells/ μ l)	0.610	0.024	0.611	0.024
Place of the program (1)	1.036	0.060	1.028	0.058
Place of the program (2)	0.800	0.072	0.795	0.071
HIV status disclosed (no vs yes)	0.684	0.054	0.679	0.053

Table 4.2: Estimates of the covariates and standard errors for the Bayesian approach for the deaths (Complete case analysis) and comparison to the frequentist one.

Upon inspection of the Tables 4.1 and 4.2, it is clear that the hazard ratios of the models of the frequentist and Bayesian approaches do not differ a lot. From the analysis we can presume that the non-informative priors that we chose for the model are suitable since the results are comparable with the complete-case analysis from chapter three. However, the standard errors of the Bayesian approach are larger than the ones of the frequentist approach. The use of non-informative priors could play a role in the standard errors. If we used certain prior beliefs about the dataset and gave specific priors, then the standard errors could be smaller. In the corresponding density and trace plots that are portrayed in the Appendix, it can be observed that the three chains have given similar results with small dis-

crepancies for the last 600 iterations. All Rhats have values smaller than 1.1 and as a result the convergence can be considered as successful. Moreover, except of the convergence diagnostics, autocorrelation diagnostics took place for the given variables and there is no evidence that there is a relationship between variable's current value and its past values.

4.3 Imputation of the missing outcomes using Bayesian logistic regression and the Weibull baseline hazard function

The second approach that will be used in chapter four, will estimate the missing causes of failure with a Bayesian multiple logistic regression model and then the received causes of failure will be used for the estimation of the hazard ratios in the survival analysis. As it has already been referred in chapter three, logistic regression is considered to be more suitable due to the fact that the patients with missing cause of failure cannot be characterized as right censorings. Patients that were in the study until the end without expressing an outcome are considered as right censorings. The steps that are needed to be done are the following:

1) From the initial dataset, isolate the patients that have as cause of failure either death or disengagement as outreached or their outcome is not known because they have not been traced from the health authorities.

2) For that sub-sample, fit the Bayesian logistic regression using the binomial distribution for the estimation of the cause of failure (0 for disengagement and 1 for death). The probability p that will be used in the binomial distribution can be extracted from the formula:

$$\text{logit}(p) = \beta_0 + \beta_1 * I(\text{sex} = \text{male}) + \beta_2 * I(\text{hivstatus} = \text{yes}) + \beta_3 * \text{Age} + \beta_4 * \text{CD4} + \beta_5 * \text{Level1} + \beta_6 * \text{Level2}, \quad (4.7)$$

3) From the Bayesian multiple logistic regression, take the estimated binary outcomes for the patients whose outcome was missing.

4) Merge the given sub-sample with the sub-sample of right censorings.



5) Fit the Bayesian cause-specific hazard model with the Weibull baseline hazard function with the same prior distributions that were used in the complete case analysis in section 4.2.

In order for the model to be fitted, 10000 iterations took place with the burn-in period to be 1000 iterations. As in the previous sections, three chains were calculated and every tenth value was kept. The results for the two causes of failure are portrayed in Tables 4.3 and 4.4.

Covariates	Hazard ratios	Standard error	Hazard ratios of the MP-PLE model	Standard error of the MP-PPLE model
Age at enrollment (per 10 years)	0.795	0.009	0.798	0.007
Gender (Female vs Male)	1.029	0.019	1.037	0.015
CD4 Cell count (per 100 cells/ μ l)	1.043	0.005	1.039	0.005
Place of the program (1)	1.149	0.021	1.164	0.014
Place of the program (2)	1.031	0.026	1.045	0.013
HIV status disclosed (no vs yes)	1.151	0.018	1.144	0.011

Table 4.3: Estimates of the covariates and standard errors for the Bayesian approach (imputation of missing values) for the outcome of disengagements and comparison to the frequentist one.

Covariates	Hazard ratios	Standard error	Hazard ratios of the MP- PLE model	Standard error of the MPPLE model
Age at enrollment (per 10 years)	1.081	0.014	1.102	0.010
Gender (Female vs Male)	1.364	0.033	1.339	0.014
CD4 Cell count (per 100 cells/ μ l)	0.665	0.015	0.676	0.013
Place of the program (1)	1.095	0.039	1.040	0.021
Place of the program (2)	0.822	0.051	0.791	0.035
HIV status disclosed (no vs yes)	0.800	0.032	0.781	0.019

Table 4.4: Estimates of the covariates and standard errors for the Bayesian approach (imputation of missing values) for the outcome of death and comparison to the frequentist one.

Upon inspection of Table 4.4 that portrays the hazard ratios and standard errors if death is the outcome of interest, small differences can be detected between the Bayesian and the frequentist model. The biggest difference can be seen in the variable of the place of the program. If two patients are compared and the only difference between them is that the one does the treatment in a Health center and the other in a district hospital, then if we trust the frequentist approach then the latter has 4% more risk than the former, while if we trust the Bayesian approach the hazard risk increases to 9.5%. In the rest hazard ratios, the differences are smaller and the discrepancy of the hazard ratios can be explained because the MPPLE approach is a semi-parametric approach, while the Bayesian one is a

fully parametric approach. Probably, if better prior distributions were selected, the results will be closer to the frequentists ones.

Moreover, if disengagements are the outcome of interest, then differences also exist but they are smaller. Also, upon inspection of the standard errors in the disengagements, standard errors are slightly smaller in the frequentist method. However, if the outcome of interest is the loss of life, the differences of the standard errors between the two frameworks are bigger. It seems that the frequentist approach gives more precise results in comparison to the Bayesian one. In conclusion, it can be said, that the Bayesian cause-specific hazard model can be characterized as an additional way to calculate the hazard ratios or confirm the frequentists methods. Last but not least, it is clear that, both with frequentist and Bayesian procedures, the imputed dataset considers CD4 cell count as a risk factor when gap in care is the outcome of interest, in comparison to the complete case analysis.

Furthermore, due to the size of the dataset and the number of iterations that were applied, the computational time can be considered as a big drawback. The laptop that was used for the thesis is a Lenovo G50-30 (2015) and the computational time was more than expected. This huge disadvantage can play a significant role in the choice of this Bayesian model, since notable approaches have given precise results in shorter time. For the visualization of the last 600 iterations for the 3 chains that were created as well as the density plots for the values, please refer to the Appendix.

Moreover, it should be noted that in this chapter, models based on the Weibull distribution were used. Other approaches could also be used, such as exponential baseline hazard functions or Cox proportional hazards models in the Bayesian framework. Furthermore, regarding the connection between the third and the fourth chapter, similar results with the Cox proportional hazards model were granted in the third chapter with the fit of the Weibull regression. As a conclusion, the results of the fourth chapter can be compared to the ones of the third. Please refer to the Appendix for the hazard ratios of the frequentist Weibull regression and the comparison to the Cox proportional hazards model.



4.4 Comparison of the models and conclusion of fourth chapter

In this chapter, the focus was to create a Bayesian model that would estimate the hazard ratios of the outcomes of interest as well as to estimate the causes of failure for the patients whose outcome is missing. The approach that was followed for the final model was the cause-specific hazards model with a Weibull baseline hazard function. In this chapter, it was very important to obtain as small standard errors as possible. Moreover, in order for comparisons to be made, a model with the misspecified causes of failure was created. In the Tables 4.5 and 4.6 below, a comparison of the hazard ratios and their standard errors of all the aforementioned approaches is portrayed.



Covariates	Hazard ratio(Standard error) complete case analysis	Hazard ratio(Standard error) Imputed approach	Hazard ratio(Standard error) Misspecified model
Age at enrollment (per 10 years)	0.741 (0.016)	0.795 (0.009)	0.832 (0.005)
Gender (Female vs Male)	1.066 (0.032)	1.029 (0.019)	1.069 (0.008)
CD4 Cell count (per 100 cells/ μ l)	0.991 (0.010)	1.043 (0.005)	1.005 (0.017)
Place of the program (1)	1.078 (0.039)	1.149 (0.021)	1.174 (0.016)
Place of the program (2)	1.050 (0.043)	1.031 (0.026)	1.005 (0.021)
HIV status disclosed (no vs yes)	1.019 (0.031)	1.151 (0.018)	1.105 (0.026)

Table 4.5: Comparison of the misspecified model with the complete case analysis and the imputed model (disengagements)

Covariates	Hazard ratio(Standard error) complete case analysis	Hazard ratio(Standard error) Imputed approach	Hazard ratio(Standard error) Misspecified model
Age at enrollment (per 10 years)	1.052 (0.023)	1.081 (0.014)	1.101 (0.024)
Gender (Female vs Male)	1.377 (0.050)	1.364 (0.033)	1.493 (0.022)
CD4 Cell count (per 100 cells/ μ l)	0.610 (0.024)	0.665 (0.015)	0.607 (0.049)
Place of the program (1)	1.036 (0.060)	1.095 (0.039)	0.863 (0.053)
Place of the program (2)	0.800 (0.072)	0.822 (0.051)	0.822 (0.057)
HIV status disclosed (no vs yes)	0.684 (0.054)	0.800 (0.032)	0.711 (0.072)

Table 4.6: Comparison of the misspecified model with the complete case analysis and the imputed model (deaths)

First of all, the standard errors for the imputed model are smaller in comparison to the one with the complete cases alone, due to the fact that the dataset that is used is bigger and imputation has taken place. Upon inspection of the misspecified model, we can observe that there are differences between that model and the imputed one. This highlights the need for estimating the outcomes of interest. If the misspecified model was the main model for our analysis, then the gender variable would indicate that men have approximately 50% higher hazard than women for example. As a result, all the attention would be paid into this particular group.



Chapter 5

Conclusion and further discussion

Statistics play an important role in health sciences and in the society in general. Thanks to the survival analysis, the estimation of mortality rates as well as factors that influence a disease is possible and as a result, the science community is capable of taking important decisions to improve health both for the individual and public health. Also, competing risks analysis provides a set of powerful statistical techniques for analyzing survival data with multiple mutually exclusive events or causes of failure. Competing risks methods are also used when the events of interest are not mutually exclusive but the interest is focused on the first coming event.

In this thesis, we described the theory of survival analysis as well as the theory that lies behind competing risks. We focus on the approaches that can help us estimate the mortality rates and the number of people that have experienced each cause of failure. The motivation of this thesis came from the IeDEA study of HIV-infected individuals in East Africa. In this study, there was a significant amount of death under-reporting. To deal with this issue, IeDEA investigators implemented a double sampling design to transform the misclassification problem due to the death under-reporting, into a missing data problem. To this end, we considered in this thesis, methods for competing risks analysis with missing causes of failure. In the frequentist framework, it was clear that the approach that uses the Maximum pseudo-partial-likelihood estimation (MPPLE) gives more precise results since the standard errors are much smaller in comparison to the other proposed methods such as multiple imputation or omitting all the missing values (complete case analysis). Also, the complete case analysis is expected to provide seriously biased



estimates.

Regarding the hazard of death, it seems that the gender of the patient is very important, since male patients have approximately a 35% higher hazard of death in comparison to women. Another very important variable is the CD4 cell count, because if these cells are decreasing, then the patients are more likely to die. These facts can give more reasons to the science community to be more proactive and take care of the most vulnerable groups of patients.

Regarding the hazard of disengagement from care, an important fact is that younger patients have a higher hazard of disengagement from care. The immaturity of the young people can be a decisive factor for the last result. Moreover it is obvious how significant is the imputation of the missing values. If the missing outcomes of interest are omitted then the CD4 cell count variable can be considered as a protective factor since the hazard ratio is lower than one. However, if the missing outcomes are imputed, either with the approach of multiple imputation or with the MPPL method, the CD4 cell count is considered as a risk factor since the hazard ratio is higher than one. The same happens in the Bayesian framework, where the complete cases analysis "views" CD4 cell count as a risk factor, while with the imputation technique CD4 cell count is considered to be protective.

The frequentist approach can give accurate results for our analysis. Similar results can also be received using Bayesian statistics. With the difference on how the missing values are considered, this approach can give precise results. For our thesis, a cause-specific hazards model with a Weibull baseline hazard function was applied and the aforementioned non-informative priors in sections 4.2 and 4.3 were used. This model is fully parametric in comparison to the semi-parametric model that was used in the MPPL method. The results have shown that the Bayesian methods can be an extra asset in the estimation of the mortality rates as well as the number of patients that have lost their lives. Furthermore, the standard errors of the Bayesian models are quite similar to the ones that have been granted in the frequentist methods and as a result precise results can be given.

Obviously in both frameworks that we have worked, if the analysis would be done using the misclassified cause of failure, the received results would be very different. This large discrepancy highlights the need of data to be recorded properly as well as the need for creating methods so that the parameters of interest can be estimated correctly. Furthermore, as it has already been mentioned many times in this thesis, if the analysis derives from incorrect data, incorrect conclusions will be drawn. One of these conclusions would be the underestimation of death, since all the not outreached patients would be considered as disengagers.

The main point of the thesis was to compare different approaches. Next steps



in this analysis could be the use of time-varying covariates or multi-state models. Also, since our dataset has variables that give insight about the location of the patients, spatial cluster analysis could be used. Spatial cluster analysis can help the scientists to identify whether the location of a cluster of patients and their neighbors' locations plays important role in the transmission of the disease or its mortality rate. The need for better understanding and knowledge about this situation, is crucial.





Appendix A

Appendix

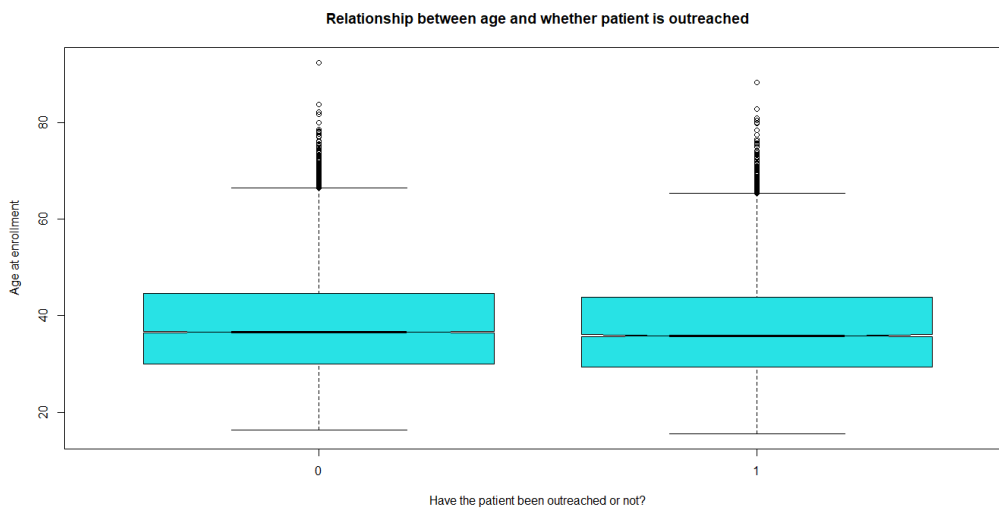


Figure A.1: Box-plot that indicates the difference in age between outreached and not outreached.



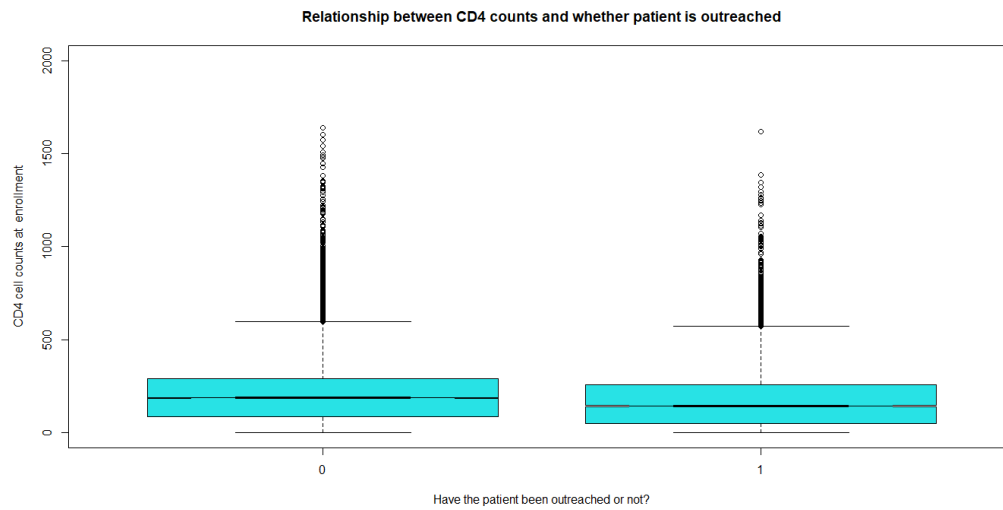


Figure A.2: Box-plot that indicates the difference in CD4 cell count between outreached and not outreached.

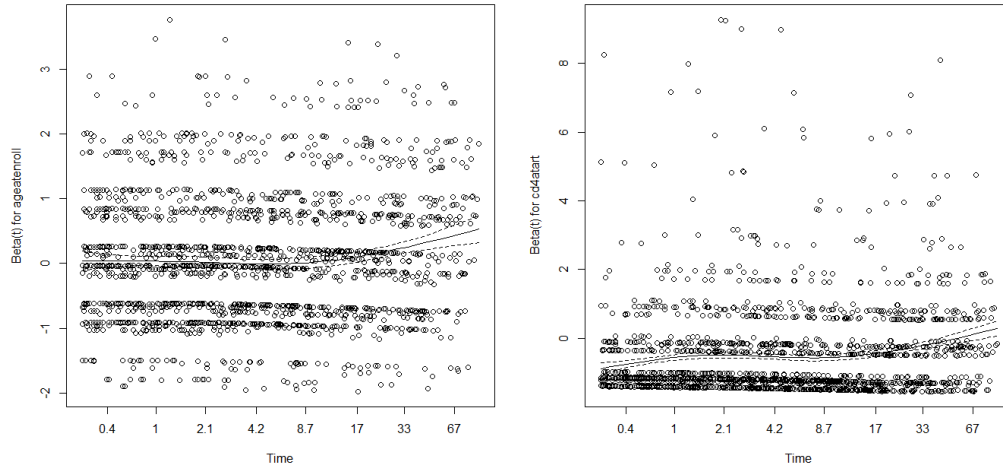


Figure A.3: Scaled Schoenfeld residuals through time for age and CD4 count for the outcome of death.

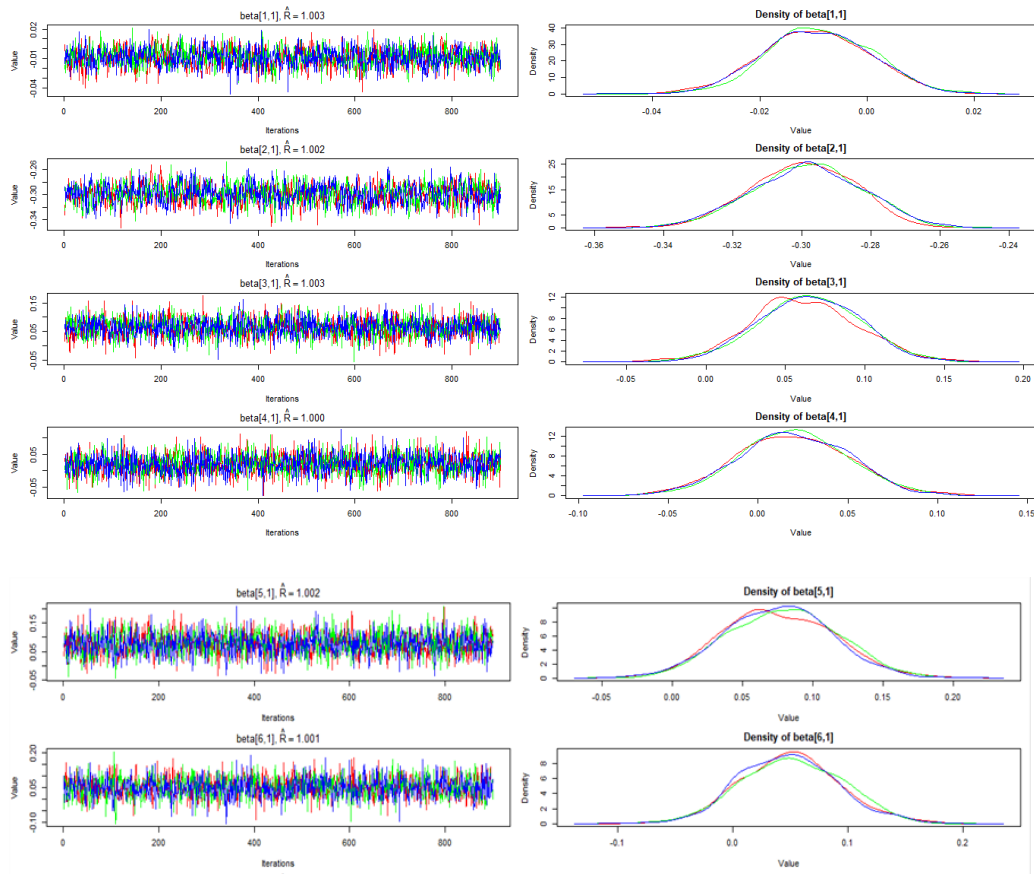


Figure A.4: Trace plots and density plots for the covariates of interest of the Bayesian complete case analysis for disengagers.



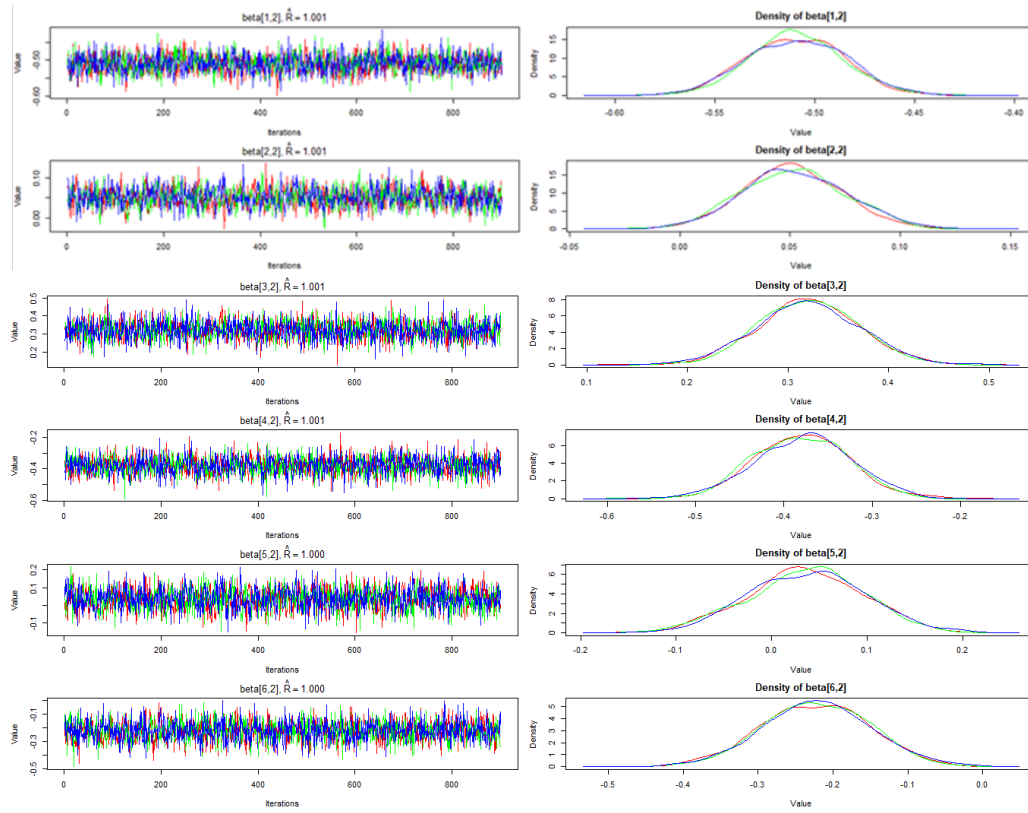


Figure A.5: Trace plots and density plots for the covariates of interest of the Bayesian complete case analysis for deaths.



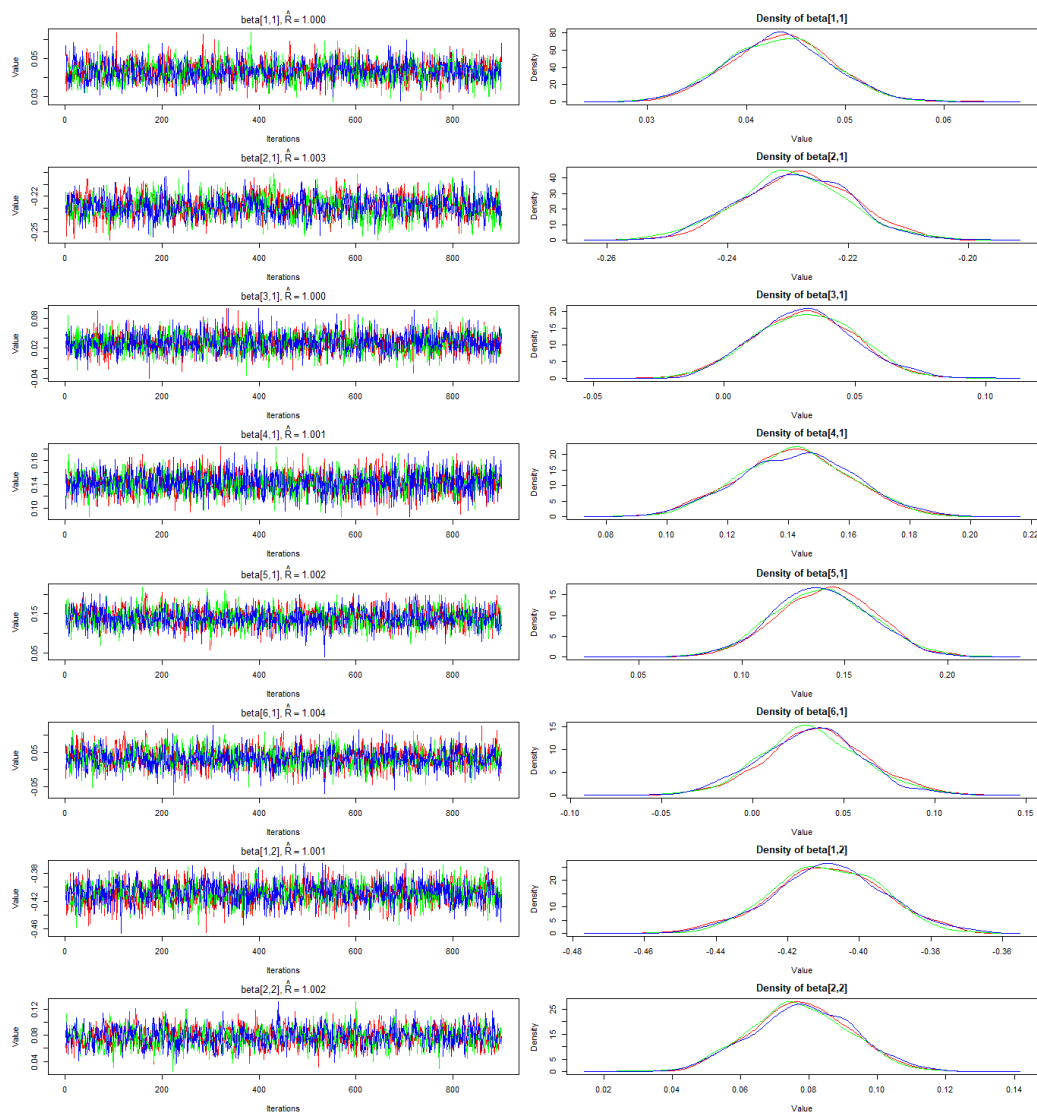


Figure A.6: Trace plots and density plots for the covariates of interest of the Bayesian imputed analysis for disengagers.



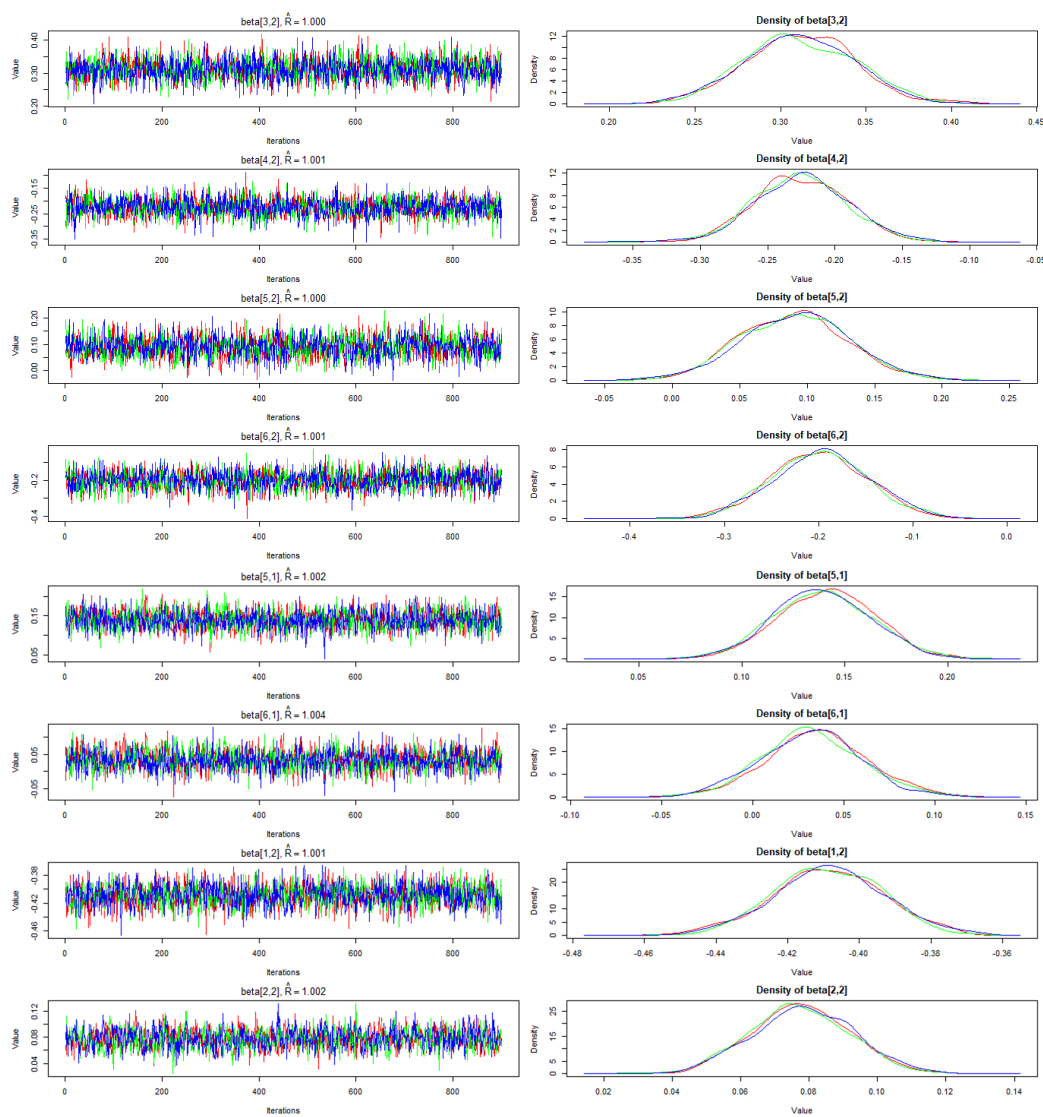


Figure A.7: Trace plots and density plots for the covariates of interest of the Bayesian imputed analysis for deaths.



Covariates	Cox hazard ratios-death	Weibull hazard ratios-death	Cox hazard ratios-disengagement	Weibull hazard ratios-disengagement
Age at enrollment (per 10 years)	1.077	1.063	0.750	0.741
Gender (Female vs Male)	1.364	1.373	1.071	1.073
CD4 Cell count (per 100 cells/ μ l)	0.611	0.605	0.984	0.990
Place of the program (1)	1.028	1.035	1.084	1.080
Place of the program (2)	0.795	0.798	1.052	1.049
HIV status disclosed (no vs yes)	0.679	0.684	1.004	1.023

Table A.1: Comparison of the hazard ratios in the Complete case analysis for the Cox proportional hazards model and Weibull regression.

Covariates	Cox hazard ratios-death	Weibull hazard ratios-death	Cox hazard ratios-disengagement	Weibull hazard ratios-disengagement
Age at enrollment (per 10 years)	1.102	1.090	0.798	0.791
Gender (Female vs Male)	1.339	1.343	1.037	1.037
CD4 Cell count (per 100 cells/ μ l)	0.676	0.674	1.039	1.042
Place of the program (1)	1.040	1.057	1.164	1.172
Place of the program (2)	0.791	0.791	1.045	1.041
HIV status disclosed (no vs yes)	0.781	0.792	1.144	1.154

Table A.2: Comparison of the hazard ratios in the MPPLE approach for the Cox proportional hazards model and Weibull regression.



Bibliography

Wenxian Zhou et al. (2022). Semiparametric marginal regression for clustered competing risks data with missing cause of failure.

Giorgos Bakoyannis et al. (2020). Semiparametric regression and risk prediction with competing risks data under missing cause of failure.

Yiannoutsos et al. (2008). Sampling-based approaches to improve estimation of mortality among patients dropouts: experience from a large PEPFAR-funded program in Western Kenya.

David Collett (2015). Modelling survival data in medical research (Third edition).

Michael T. Koller et al. (2010). Competing risks and the clinical community : irrelevance or ignorance?

Danilo Alvares et al. (2020). Bayesian survival analysis with BUGS.

Zhongheng Zhang (2017). Survival analysis in the presence of competing risks.

Ioannis Ntzoufras (2009). Bayesian modeling using WinBugs.

K.K. Vidya Vijayan et al. (2017). Pathophysiology of CD4+ T-Cell Depletion in HIV-1 and HIV-2 Infections.

Giorgos Bakoyannis and Giota Touloumi (2010). Practical methods for competing risks data: A review

Per Kragh Andersen (2011). Competing risks in epidemiology: possibilities and pitfalls.

Xanthi Pedeli (2021). Biostatistics- Survival analysis.

Gao and Tsiatis (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure.

Peter C Austin (2022). Estimation of the absolute risk of cardiovascular disease and other events: Issues with the use of multiple Fine-Gray subdistribution Hazard models.



Craiu and Duchesne (2004). Using EM and data augmentation for the competing risks model.

Gray (1988). A class of K-Sample tests for comparing the cumulative incidence of a competing risk.

Robins and Rotnitzky (1992). AIDS Epidemiology - Methodological Issues

Lancy et al. (2016). Kaplan-Meier survival analysis overestimates cumulative incidence of health-related events in competing risk settings: a meta-analysis.

