ΠΜΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

ΑΠΟΣΤΟΛΙΔΗΣ ΑΠΟΣΤΟΛΟΣ

A simulation study of the Bayesian Weibull competing risk model with missing cause of failure.

ΑΘΗΝΑ, 2023

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη

ΒΙΟΣΤΑΤΙΣΤΙΚΗ

που απονέμει η Ιατρική Σχολή και το Τμήμα Μαθηματικών του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε την…….. από την εξεταστική επιτροπή:

ΟΝΟΜΑΤΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ	ΥΠΟΓΡΑΦΗ
1.		
2.		
3.		

Thank God and all the people that supported me.

Contents

1 Introduction

The main purpose of this master thesis is to describe the Bayesian Weibull competing risks model with missing event types and then compare it with an existing model in which its coefficients are estimated via the maximum pseudo partial likelihood method. In the second and third chapters, the essential theory, for the description of the model, is written. In the second chapter, the theory about the competing risks and in general the survival analysis is presented. Particularly, some elementary theory about the survival analysis is described and then the two different types of modeling the competing risks are defined. In the same chapter, some basic theory about the missing values is presented, and how each missing scenario can be solved. In the end, there are presented some methods to tackle and model the competing risks with missing cause of failure scenarios. In the next chapter, the third one, the Bayesian theory, the Weibull survival model, and their combination are thoroughly explained. The Bayesian Weibull Competing risks with missing cause of failure is a combination of four theories, the Bayesian, the competing risks, the Weibull survival model, and a Bayesian methodology of imputing the missing values. Therefore, the first 2 main chapters try to describe and analyze the theory behind this complex and complicated model.

In the fourth and fifth chapters, the methodology and the results are presented. Particularly, in the fourth chapter, the desirable two models are meticulously described. First, the Bayesian Weibull competing risk with the missing cause of failure model and then the other model which uses the maximum pseudo partial likelihood. Those two methods, especially the target method are both theoretically and practically explained. The algorithms for both models are precisely described and the code for the main method is given in the Appendix. Furthermore, the theory behind the data simulation is given, and the different scenarios are derived from it. Also, descriptive statistics and some analytic graphs are given in order to somehow enlighten the structure of the data. In the fifth chapter, the parameters of the Bayesian algorithm are derived from prior knowledge (knowledge from another study) and from the trial-and-error estimation method. Next, the results and the evaluation metrics are described. The results are basically given in tables and graphs. In addition, the convergence is measured by different metrics such as the scale reduction factor and the bias (because the results are known). In the end, some sensitivity analyses are conducted only for the first scenario because of the computer and time capabilities.

Last but not least, the discussion is the essence of this master thesis because all the advantages and disadvantages of the study are presented there. Also, the advantages and disadvantages of the models are thoroughly explained. In the end, some final thoughts or some study proposals are given. After that, the summary, both in English and Greek, the reference, and the appendix which has the code of simulation, and the estimation are given. The code is written in R (version 4.3.0).

2 Competing risks with missing event types

2.1 Competing Risks

2.1.1 Survival Analysis

Competing risk usually arises in survival analysis when the event of interest cannot occur because another event has already occurred[1,2]. In other words when they are more than one event and usually, scientists are interested in one of them. For example, in cancer research, the event of interest can be death from cancer and the other event can be death from another reason [1]. When estimating the cumulative incidence analysts must adjust for the multiple events and do not use the naïve approach of Kaplan Meier [1]. That is to say that if the other events are converted to censor observations, an analyst can estimate the Kaplan Meier estimator and that estimator is biased upwards regardless of the event's relation[1]. So, researchers always should adjust their analysis when competing risks are present. There are two ways to adapt the analysis to competing risks first the researchers can choose the cause-specific hazard function with the purpose of estimating the hazard function and the other is modeling the cumulative incidence function using the subdistribution hazard function [1,2,3]. Both ways are modeling the effect of covariates the former is trying to model the effect of covariates on each hazard and the other the effect of covariates on cumulative incidence function[1,2,3]. As it is said the first is better to address etiological questions and the other is to estimate the clinical prognosis of a patient. Particularly, the former is to estimate the effect of covariates on the rate of occurrence of the event when the person is event-free. The latter allows the estimation of how the covariates affect the absolute risk of the outcome over time [1].

It is assumed that T donates the time to event in other words the time from baseline time until the occurrence of the desirable event. When the competing risk is absent the survival function $S(t)$ is the probability of the outcome occurrence after time t can be easily estimated by Kaplan Meier method. So the $S(t)$ is described as the probability of certain subject survives at least times t or in mathematical way as $S(t) = P(T > t)$ or $1 - S(t) = F(t)$ where $F(t)$ is cumulative function $P(T \leq t)$. As a result $S(0)$ equals to one and $S(t)$ as t tends to infinite equals to zero. Nevertheless, the final property is not applied to every problem because the probability of survival under administrative times is below one. In the competing risk, the Cumulative Incidence Function as $1 - S(t)$ is the probability of the survival from all distinct outcomes and allows researchers to estimate the incidence of occurrence of an outcome after adjusting for competing risks [1]. Consequently , it is defined that the cumulative incidence of the kth event is the $CIF_k = P(T \leq t,\ D = k)$ where D is the variable which donates the particular outcome [1]. In other words, CIF_k is the probability of the occurrence of the kth event before another event takes place [1]. Another desirable property is that the sum of CIF_k is the probability of any events occurrence before time t. One relatively different property is that CIF_k will not tend to unity as time goes to infinity because another event might happen before the event of interest or just in the end one event can occur. In the absence of competing risk, the hazard

function is defined as $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t+\Delta t | T \geq t)}{\Delta t}$ [1]. In other words, the hazard function which is function of time describes the instantaneous rate of occurrence of a desirable outcome when the subject is event free. The well known, Cox proportional hazard regression model connects the hazard function with the covariates. More particularly, when the competing risks are absent the Cox proportional hazard regression can be described as

$$\log h(t) = \log h_0(t) + X\beta$$

where $h_0(t)$ [2] is the baseline hazard function typically when covariates are zero and X is the design matrix multiplied with the coefficients β. If an exponential transformation is used the former relation is written as $h(t) = h_0(t)e^{X\beta}$. The coefficients are basically equal to log hazard ratio and they are interpretated as the relative change of the log hazard function when the relative covariates are changed by one unit [2]. If the relation $S(t) = e^{\int h(t)dt}$ is used, $S(t)$ can be written as $S(t) = S_0(t)^{\exp(X\beta)}$ [2] where $S(t)$ is the survival function and $S_0(t)$ is the baseline survival function (covariates equal to zero) under the absence of competing risk. That is to say that the inference about hazard rate is related to the survival function. A more particularly positive coefficient means a bigger hazard ratio, but a lower survival and negative coefficient means a lower hazard ratio as a result of bigger survival [2]. Finally, every researcher should check for the assumption that the hazard ratio is independent of time, or the proportionality property of the hazard function is valid. In other words, the hazard ratio is not a function of time as a result it is invariable as time moves on. This is an important property that analysts must check every time. If this property is not valid then the researcher can proceed to other models like parametric ones or can tune the Cox model and a time covariate in the equation[4].

Now in the case of the competing risks which is that there are multiple and distinct events supposing one of them can happen in the study. An individual can experience one of many events but only one and as it has been mentioned there are two different hazard functions or types of modeling to adjust to the competing risk framework. The first one is the cause-specific hazard function which practically models the hazard function for each outcome when the subject is event-free and the second one is the subdistribution hazard function which tries to model the cumulative incidence function[1,2,3]. So, the formula for the cause-specific hazard function is

$$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t+\Delta t, D=k | T \geq t)}{\Delta t}$$

and it is recognized that the difference from the not competing risk hazard function is that this so-called cause specific hazard function adjust for specific kth event in the probability or in the numerator of the limit [1,2,3]. Again, cause specific hazard function is the instantaneous rate of occurrence of a kth specific outcome in an event free subject (a subject that has not experience any event). Consequently, analysts can relate the cause specific hazard function of a specific event to covariates through coefficients. In other word cause specific hazard function is frequently used for etiological reasons in opposite to subdistribution

hazard function. In the opposite side of things, the subdistribution hazard function has a little more complicated formula which is

$$h_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t+\Delta t, D=k | T > t \cup (T < t \cap K \neq k))}{\Delta t}$$

and it was introduced by Fine and Gray [5]. It is noticeable that an extra term is introduced in the conditional probability , so the condition term is the union of an event free time and an outcome time of a different event. In other words is the instantaneous risk of failure in subject which have not experienced the kth outcome [5,1,2,3]. The main difference between the hazards is that cause specific hazard function is estimated for event free subjects and subdistribution hazard is estimated for a kth event free subject. Subdistribution hazard relates the effect of the covariates to the cumulative incidence function therefore the predicted CIF_k can be estimated for a certain subject and that is why subdistribution hazard is better for prognosis purposes of a specific event in a particular subject [1,2]. It is preferred that analysts should estimate both hazards when both hazards can be estimated and when the scientific question is ambiguous.

2.1.2 Cause Specific Hazard function

It is frequently phenomenon that there are more than one distinct event of interest and one way to adjust for this problem is to use cause specific hazard function. The cause specific hazard function is perfect for answering etiological questions and it donates the instantaneous rate of occurrence of a specific outcome on free event subject. It is defined as

$$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t+\Delta t, D=k | T \geq t)}{\Delta t}$$

so, it same as non-competing risk hazard with the difference that in probability there is an indicator which shows for which event the hazard rate is calculated. The cumulative probability of the realization of an outcome before time t when all other outcomes are possible to happen is known as cumulative incidence function and it is defined as

$$F_k(t) = P(T \leq t,\ D = k) = \int_0^t \mathrm{h}_k^{cs}(u)\, e^{-\int_0^u \Sigma_{j=1}^K \mathrm{h}_j^{cs}(v)dv} du$$

where K is the number of the possible events and this formula is only for cause specific hazard function [3]. So, under the assumption of the independent right censoring which means that the censoring time is independent of the event times and cause of failure, the likelihood function can be described as

$$L = \prod_{i=1}^{n} \prod_{k=1}^{K} [h_k^{cs}(t_i, x_i)]^{d_{ki}}\, e^{-\int_0^{t_i} h_k^{cs}(v, x_i)dv}$$

where n is the total number of observations, t_i is the time until the realization of an outcome d_{ki} is an indicater which is equal to 1 if ith subject has experienced the kth event and 0 elsewhere and x_i is a design matrix[3]. So, after utilizing the property of permutation the likelihood can also be written as

$$L = \prod_{k=1}^{K} \prod_{i=1}^{n} [h_k^{cs}(t_i, x_i)]^{d_{ki}} \, e^{-\int_0^{t_i} h_k^{cs}(v, x_i) dv}$$

which is the product of the likelihoods for each event[3]. In other words, is the total product of each event likelihood

$$L = \prod_{k=1}^{K} L_k$$

where $L_k = \prod_{i=1}^{n} [h_k^{cs}(t_i, x_i)]^{d_{ki}} \, e^{-\int_0^{t_i} h_k^{cs}(v, x_i) dv} = \prod_{i=1}^{n} [h_k^{cs}(t_i, x_i)]^{d_{ki}} \, S_k^{cs}(t_i, x_i)$ which is the likelihood if it is presumed that the other events but k are censored and $S_k^{cs}(t_i, x_i)$ is the typical non-competing risk survival function for kth event if observations with different events are treated as censored[3]. In plain words it is vital to procced to K analysis one for each event and treat the other events as censored observations assuming that the same predictors and the same methods are used. In other words if a competing survival analysis is conducting with two distinct event, the analyst can separate the competing analysis to two plain survival analysis replacing the other event with censoring but the data and the covariates are the same. So, because of the previous beneficial property in independent right censoring the cause specific hazard is sometimes preferred than the subdistribution one. In conclusion the cause specific hazard function $h_k^{cs}(t_i, x_i)$ can be modeled as a normal $h(t_i, x_i)$ just by treating the non k events as censored ones. Particularly a Cox model can be used in order to analyze a competing risk data set just by conducting Cox modelling for each event, thus the requested cause specific hazard function for a specific kth event is

$$h_k^{cs}(t_i, x_i) = h_{k0}^{cs}(t_i, x_i) e^{\beta_k x_i}$$

Where $h_{k0}^{cs}(t_i, x_i)$ is the baseline cause specific hazard function and $\beta x_i = \sum_{p=1}^{P} \beta_{pk} x_{pi}$ where β_{pk} is the coefficient for X_p variable $p = 1, 2, \ldots, P$ are the predicted variables for the k event[3]. For each analysis P coefficietns are estimated by either maximizing the likelihood or by using Bayesian statistics. In the end $P *$ K coefficients are estimated (P for each analysis). In each analysis, the same covariates must be inserted in the model and the final model evaluation is done by evaluating each model separately. Apart from using a semi-parametric proportional Cox model, an analyst can use a parametric one like an accelerated time failure model by giving the baseline hazard function a formula. So, if a researcher is suspicious that the proportional hazard might not be valid then they can proceed to accelerate time failure models which those models have various types like Weibull, log-logistic, gamma, and more each of them models hazard functions differently for example Weibull hazard function is monotonous in opposition to log-logistic which is not monotonous[4]. More information about those models is mentioned in the next chapter. So, the proportionality property for each event must be meticulously examined and that can be succeeded through various tests and graphical diagnostics using for example Schoenfeld residuals [4]. If the proportional hazard assumption is not valid apart from using parametric models an analyst can insert the time to predicted variables but this is easier said than done or a researcher can use stratified Cox models for the variable whose

proportional hazards assumption is not true. Finally, the variable selection is conducted using likelihood tests by means of using the property that the total likelihood of a model competing risk model is the product of each likelihood when the other events are censored. Apart from the hazard modeling, a researcher can model the cumulative incidence function utilizing subdistribution models.

2.1.3 Subdistribution hazard function.

Another approach to model the competing risk problem rather than cause specific hazard function is using subdistribution hazard function [5]. Subdistribution hazard modelling was introduced by Fine and Gray [5] and this type of model also is called as CIF regression model or Fine and Gray hazard or model. Calling Subdistribution hazard model as CIF regression model, an individual can understand the strong connection between subdistribution hazard and cumulative incidence $F_k(t) = P(T \leq t,\ D = k)$. Particularly

$$h_k^{sub}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t, D = k | T > t\ \cup (T < t\ \cap K\ \neq k))}{\Delta t}$$

$$= \frac{[\frac{dF_k(t)}{dt}]}{1 - F_k(t)} = -\frac{d}{dt}\log\left(1 - F_k(t)\right)$$

and this can be generalized with the introduction of covariates x

$$h_k^{sub}(t, x) = -\frac{d}{dt}\log\left(1 - F_k(t, x)\right)$$

As it is mentioned the subdistribution hazard is linked to the cumulative incidence in opposition to cause specific hazard function. Consequently, inference can be made for the cumulative incidence using the subdistribution hazard function[5,1,2,3]. Nevertheless, a researcher should interpret the results with care because interpreting the effect of covariates in the subdistribution hazard model is a little unnatural because the Fine and Gray hazard is the incidence of a covariate in a subject that is event-free (from the event of interest k) and is not event free from other outcomes. For example, if one event is death and the other is cure the estimation of subdistribution hazards for death is made by subjects who are not either dead or cured or they just are cured[1,2]. So, the interpretation is done either by interpreting the rate of hazard in event-free subjects or by connecting the cumulative incidence function to subdistribution hazards.

Fine and Gray [5] introduced a semiparametric proportional hazard model for the desirable event :

$$h_k^{sub}(t, x) = h_{0k}^{sub}(t, x)\, e^{x\beta_k}$$

Where $h_{0k}^{sub}(t, x)$ is the base line subdistribution hazard and x are the covariates with β_k coefficients for the k event. The cumulative incidence using the subdistribution hazard is described as

$$CIF_k = F_k(t,x) = 1 - e^{-\int_0^t h_k^{sub}(u,x)du} = 1 - e^{-\int_0^t h_{0k}^{sub}(u,x)\, e^{x\beta_k} du}$$
$$= 1 - e^{-\int_0^t h_{0k}^{sub}(u,x)\, e^{x\beta_k} du} = 1 - \ [1 - F_k(t,0)]^{e^{x\beta_k}}$$

resulting in

$$1 - F_k(t,x) = [1 - F_k(t,0)]^{e^{x\beta_k}}$$

where $F_k(t,0)$ is the base line cumulative incidence for the kth event[5,2,3]. As it is observable the coefficients β_k are present in the hazard equation and in the incidence function equation. This means that researcher can use coefficients β_k to interpret the CIF_k but it is not clearly what type of interpretation can be achieved because the previous is not very easy to interpret[2]. Someone can use log transformation $\log[1 - F_k(t,x)] = e^{x\beta_k} + \log[1 - F_k(t,0)]$ and then interpret that the logarithm of rate ratio of one minus cumulative incidence function when covariates are x to one minus cumulative incidence function when covariates are 0 is $e^{x\beta_k}$ [2]. Mathematically this is written as $\log[\frac{1-F_k(t,x)}{1-F_k(t,0)}] = e^{x\beta_k}$ and utilizing the equation with hazard and the baseline hazard $\frac{h_k^{sub}(t,x)}{h_{0k}^{sub}(t,x)} = e^{x\beta_k}$, it is derived to the fact that $\log[\frac{1-F_k(t,x)}{1-F_k(t,0)}] = \frac{h_k^{sub}(t,x)}{h_{0k}^{sub}(t,x)} = e^{x\beta_k}$. So, if the rate of hazards is bigger than one then the one minus cumulative incidence is bigger than its one minus baseline cumulative incidence. Therefore, the direction is the same for the hazard ratio and one minus cumulative incidence ratio but the quantification is not the same[2]. So, there is one relation between hazards and cumulative incidence and that is the influencing and desirable power of subdistribution hazard modeling. Practically Fine and Gray modeling can give us information about the incidence or the survival prognosis of a patient thus it can solve problems like what is the probability of living at most T times without the event k occurring[2]. In general, it is preferable to interpret the results using cumulative incidence because it does not involve the concept of an event-free subject or event-kth-free subject which sometimes seems irrational as explained above. In opposition to subdistribution hazards specific hazards cannot directly connect to cumulative incidence, in other words, there is not one to one connection between the cause-specific hazard function and cumulative incidence.

Estimating the coefficients in CIF regression modeling is considered hard relative to causing specific hazard modeling. This is not because the actual estimation method is cumbersome or complex but because of the censoring[5,3]. First of all, in most problems independent right censoring is rightly assumed and the methods for this type of censoring are elementary ones for example for non-competing risk survival problems or competing risk problems utilizing cause-specific hazard modelling, but this is not applied to the sub-distribution hazard models. To begin with, the simplest scenario is to not have censored observations because each subject has experienced an event, according to the Fine and Gray modeling [5,3]

$$h_k^{sub}(t,x) = \ h_{0k}^{sub}(t,x)\ e^{x\beta_k}$$

It is valid to estimate each vector (β_k) of coefficient just by maximizing the likelihood of the specific event kth. So, the requested likelihood for kth event which has to be maximized is

$$L_k = \prod_{i=1}^{n} [\frac{e^{x_i \beta_\kappa}}{\sum_{j \in R_i} e^{x_j \beta_\kappa}}]^{\delta_{ik}}$$

where n is the total number of observations, $x_i \beta_\kappa = \sum_{p=1}^{P} x_{ip}\, \beta_{kp}$ where P is the number of covariates, β_{kp} is the coefficient for X_p variable in the kth event, x_{ip} is the value of ith individual in the X_p variable $p = 1, 2, \ldots, P$, δ_{ik} is an indicator which is 1 when the ith subject has experience the kth event and 0 elsewhere and finally $R_i = \{ j : T_j \geq T_i \cup [(T_j \leq T_i) \cap (\delta_{jk} \neq 1)] \}$ [5]. This likelihood is called partial likelihood, it is maximized by the frequent tools (take the logarithm and the rest) and it has all the good properties of likelihood like asymptotic normality and consistency of estimators. Nevertheless, when censoring is introduced to the framework, the situation and the solution to the problem are partially changed [5,3]. Assuming that there is an administrative censoring, which is the censoring due to the end of the study, one for a specific event let's say k can treat the non-censored times of the other non-kth events as censored ones and then proceed to the normal cox analysis[5,3]. This is the same as the cause–specific hazard. In other words, one can conduct a k analysis for each outcome and each of them can mark the observation that experiences another event as censored. Then can proceed to the normal Cox analysis. When it comes to prominent and most probable independent right censoring the concept is altered.

The independent right censoring is an issue because practically the censoring time for each event is needed to estimate the coefficients for each event. In other words, if an analyst is interested in the kth event, they must have the censored time for each observation that has experienced one of the other competing risks to estimate the particular kth coefficients. So that desirable censored time is unknown because the observations are not censored. After all, they have experienced another outcome. For example, if the event of interest is the first then for each observation that has not experienced the first event their censored time to the first event is needed but those observations have experienced other events, so their requested censored time is missing. For the problem of the independent right censoring an analyst can use the inverse probability of censoring weighting which by and large assigns every observation a weight [5,3]. Furthermore, there are more ways to deal with the problem like utilizing multiple imputations[6]. The basic concept is that the unknown censoring times are imputed by random values from a specific censoring distribution which is conditional based on the occurrence of censoring which is done after the realization of one or another competing events. This specific and desirable distribution is estimated by the Kaplan-Meier method by replacing the censored observation as an outcome and the actual event times as censored [6]. As mentioned, another meaningful way to adjust for independent right censoring is inverse probability of censoring weighting[5,3]. This technique is very popular because it is generally very understandable because an analyst just assigns weight to observations and because it is also used to solve other problems like missing values. Particularly,

assuming that there are n observations and let T, δ, ε and X are the observed time, the event indicator, the cause of failure and final the covariates. The set ε is the total distinct types of outcomes les say $\varepsilon = 1, 2, 3 \ldots, P$ and in this specific example the coefficients for the first event are requested to be calculated. Every observation ith has unique $\delta_i = 0$ if the observation is censored and 1 if it is not,ε_i equals to the type of the event, T_i the value of T column and their covariates X_i. Now because the coefficients for the first are needed to estimate if $\varepsilon_i = 1$ then let $N_i(t) = I(T_i \leq t)$ and $Y_i(t) = I(T_i \geq t)$ and for $\varepsilon_i \neq 1$ then $N_i(t) = 0$ and $Y_i(t) = 1$. Let $r_i(t)$ a function which If δ_i is 0 then $r_i(t) = 1$ when $t \leq T_i$ and 0 otherwise and if $\delta_i = 1$ then $r_i(t) = 1$ and weight $w_i(t) = r_i(t)\frac{G(t)}{G(T_i)}$ where $G(t)$ is the estimation of Kaplan – Meire survival function of censoring variable and it calculated using $\{T_i, 1 - \delta_i, i = 1,2,3,..n\}$.So if $t \leq T_i$ then $w_i(t) = 1$ and if $t > T_i$ and $\delta_i = 0$ then $w_i(t) = 0$ else

$(\delta_i = 1 \; and \; t > T_i)$ $w_i(t) = \frac{G(t)}{G(T_i)}$. Therefore, to estimate the β_1 coefficients for the first event a pseudo-likelihood L_1 are to be maximized

$$L_1 = \prod_{i=1}^{n} [\frac{e^{\beta 1 x_1}}{\sum_{j=1}^{n} Y_j(T_i) w_j(T_i) e^{\beta 1 x_1}}]^{I(\delta_i \varepsilon_i = 1)}$$

The above algorithm is applied for every event to find the coefficients for each outcome relatively. The disadvantage of the inverse probability of censoring weights is that it requires a special computing function in a statistical programming language to maximize the pseudo–likelihood. Nevertheless, the multiple imputation method can be achieved just with the normal survival Cox function utilizing the regeneration of multiple datasets[6,3]. Another method to account for right independent censoring is based on a weighted product–limit estimator[3]. Similar techniques like the above are frequently used in survival and complete data analysis.

To summarize there are two ways to deal with multiple events in survival analysis the first one is modeling the cause-specific hazard function which is used for etiological reasons and the other is subdistribution hazard which is used for a subject prognosis. The first models each event hazard independently and the second connects the cumulative incidence function with the specific subdistribution hazard. The risk set of the first method is subjects that are event-free at a specific time and the second one uses event-free subjects and subjects that have experienced another event. The former does not have a one-to-one connection, but the latter does. Also, the first is relatively easy to model in independent right censoring and the second one needs special care. In terms of interpretation, the first is the ratio of hazards in event-free populations and the second one is the ratio of hazards in event-free or not from another event population.

2.2 Missing values

2.2.1 Introduction

The missing values problem arises almost in every data analysis and it is a frequent phenomenon that needs an elegant approach. An analyst is requested to handle missing values and it is in their responsibility to solve this problem. The general approach to handle this issue is to first identify the type of missing values and then proceed to a deliberate approach. Sometimes those issues need multidisciplinary endeavor to mitigate the problem. There are three types of missingness first and easiest type to handle is Missing Complete At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR)[7,8,9,10,11]. Finally, it is important to identify which variable has missing values and this has to do with the role of variable with missing values. In other words, there are different ways to deal with that problem when the variable is dependent, or independent or there are missing values both in dependent and independent variables[10,11]. Finally, there are some data missing patterns like univariate patterns, unit response patterns, monotone patterns, general patterns, planned missing patterns, and latent variable patterns [10,11]. The univariate pattern appears when there are missing values in one variable. A unit response pattern is when there missing values in some variables in the same observations (rows). The monotone pattern is associated with longitudinal studies when participants have started to drop out and never return. A general pattern is when in random variables values of random observations are missing. In other words, there is no specific trend related to the missing values. This type of pattern missingness is the most common. The next planned missing pattern is a pattern that exists in questionnaires, particularly in survey studies analysts can give questionnaires in which some variables are missing in some participants to lift the respondent burden. Finally, latent variable pattern is when a variable is completely missing, and it is related to latent variable analysis.

To begin with, when the data are missing completely at random(MCAR) means that the missingness is independent of observed and unobserved data [8, 10,11]. In other words, there is no systematic difference between the observation with all the data and the observations with missing data.Frequently this specific missing problem derives from unwanted and random errors for example some data were lost because of a machine error, because of some typos or they are missing due to another unknown parameter. For instance, an error in data entry, an administrative random error (missing values due to the illness of the interviewer), and due to random events like doctor appointments. This type of error just reduces the power of the analysis and it is not the reason for introducing bias because MCAR data are a random sample of the full dataset. More often than not the assumption of MCAR is unrealistic as a result a researcher rarely assumes MCAR type of missingness [8,10,11]. One way to check this type of missingness is to run an independent t-test with the dependent variable being the missing indicator one if it is missing and zero if not and with the other variables[11]. In those t-tests and if the MCAR assumption is valid the p-values of the covariates have to be relatively large close to one but again it is better not to take the risk and

assume MCAR even if the results of general linear regression are indicating the opposite. Mathematically the MCAR probability is described as

$$P(R \,/\, Y_{obs}, Y_{miss}, \varphi) \;= P(R \,/\, \varphi)$$

Where R is 1 when the value is observed and zero when is missing, Y_{obs} are the observed data, Y_{miss} are the unobserved and φ is a set of unmeasured variables. Therefore, the probability of missingness is independent of the observed and unobserved scores or values of the data but the probability of missingness still is related to unknown parameters [10,11]. In practice, this assumption is relatively easy to solve and account for. There are numerous methods that can tackle this issue and as it is mentioned the only loss is the reduction of power which in a large sample might be meaningless. If the researcher presumes this type of missing value, it is vital to proceed with care and have strong and significant evidence that this assumption is true.

The data are missing at random (MAR) when basically the missingness is related to the observed data [10,11]. In other words, the probability of missingness corresponds to other observed data in the study but it is independent of unobserved data which if they were collected then the data would have been complete. Another fact that a lot of people misunderstand is that the MAR missingness is practically not random and it is basically related to other observed variables in in the study[8,10,11]. Nevertheless, the random term is derived from the fact that the missingness is random in the group which is correlated with the missing values. For example, let's suppose that there is a study that has two variables the first variable is one kind of score and the other is age grouped into 3 groups young, middle-aged, and the elderly. The score variable has missing values, and it is observed that in the older group, there are significantly more missing values than the other groups, so the missingness is correlated with age. Therefore, in the older group, the missingness is random. In general, there are systematic differences between the observations with observed values and unobserved values. Mathematically the probability of missingness is

$$P(R \,/\, Y_{obs}, Y_{miss}, \varphi) \;= P(R \,/\, Y_{obs}, \varphi)$$

Thus, it describes that the reason for the missingness is independent of the actual missing values in our dataset. Nevertheless, it might be related to other variables that are not gathered, or are unknown in bibliographic references. The missing data have a bigger probability of being MAR than MCAR. It is a more realistic assumption, and it is not as restrictive as MCAR[8,10,11]. In most problems a MAR assumption is valid, and, in the end, it is chosen in most cases. Again one can implement various univariate t-tests to find out which variables are correlated with the targeted missing variable[10,11].In that case, some of those tests should have a relatively small p-value which indicates that the relation between the two variables is a significant one. The MAR problem has many solutions according to the data structure or the type of missing variable (dependent, independent, or both) but a bias can be introduced if the analyst deletes the missing values. Therefore, it is vital to employ a sophisticated and appropriate missing values model to tackle the issue perfectly.

Missing Not at Random (MNAR) is a missing type in which the probability of missingness is related to unobserved which hypothetically should have been collected[10,11]. In other words, if a variable X has missing values under the MNAR assumption the missingness of X is correlated with its own missing values. For instance, a study related to obesity occurs and some people do not answer the question about their weight because they are obese or just slim. So missing values are generated because of the unobserved values of the variable. The MNAR assumption is very hard to deal with and there are just a few methods that can have mediocre results there are other methods that have better results but the assumption that those methods needed are very restricted and unrealistic[8,10,11]. In addition, the MNAR assumption is not probable to test based on the data; as a result, an analyst can only theoretically presume that the missing assumption is MNAR. It is a problem that scientists try to find methods to tackle it. A lot of times sensitivity analyses are carried out in order to compare the different missing values algorithms, or another practice is that the analyst tries to find external information about the missingness.

2.2.2 Deletion Methods

There are two popular and traditional deletion methods, the first is complete-case analysis and the second one is pairwise deletion[7,8,10,11]. The complete-case analysis deletes all the observations that have at least one missing value are deleted from the analysis. Consequently, the remaining data are without missing values and that is the only advantage of this method. On the opposite side due to the deletion of observation, this surely leads to power reduction in tests. In other words, sometimes the whole part of the dataset might be removed as a result the power of various tests or methods is decreased, this phenomenon happens especially in datasets with a large percentage of missing data or variables. Except for unavoidable power reduction, an MCAR assumption is needed for the analysis result to be without bias. In most cases, the MCAR assumption is unrealistic resulting in the introduction of unwanted and certain biases. Conclusively the advantage of this method is that the final dataset is complete at the expense of power reduction and a certain bias if MCAR assumption does not hold. The pairwise deletion is partially like the complete case analysis but with one important difference, the deletion of missing observations is conducted separately in each step of the analysis. For example, the means of several variables are requested, in complete case analysis all the observations with missing values are deleted but in the pairwise deletion the reduction of observation happens only for the specific test or for the calculation of several statistics. For the computation of the first mean, only the missing observations of the first variable are deleted, and they are deleted only for that purpose and then after the deletion, the data are returned to the initial situation, and this goes on for every mean. This strategy is an improvement related to the first algorithm. Because only the necessary observations are deleted to conduct one specific test; as a result, the power of the test is not reduced like the complete case analysis. Nevertheless, this partial deletion suffers from the same bias relative issue. If the missingness does not fall under the category of MCAR the pairwise deletion results are biased.

2.2.3 Single Imputation Methods

Single Imputation Methods are methods that impute the missing value with an estimated value from the data. This procedure happens only once and there are many distinct ways to achieve this type of imputation. There are several methods for single imputation some of them are mean and median imputation, last observation carried forward, regression imputation, k nearest neighbor imputation algorithm stochastic regression imputation, and more[12,7,13,8, 10,11]. Those methods are trying to preserve the data structure and generate unbiased estimates. In the end, the result of the single imputation method is a complete dataset. The mean and the median imputation is the easiest of all, they impute the missing value with the mean or the median of the variable[7,8,10,11]. The problem with those two methods is that they do not preserve the data structure or, in other words, they change the correlation. For example, let's assume a scatterplot of two variables X and Y both variables have missing values, and both are correlated. The imputed values are steady for both variables resulting in the reduction of correlation because the sub imputed sample is steady, and it correlates zero. Under the assumption of MAR missingness, the mean imputation produces a biased estimation of the mean. The last observation carried forward (LOCF) [7] method is frequently used in time series or longitudinal studies. This method imputes the missing value in a specific time with the previous known values of this time. It is a very easy and understandable method that is carried out, especially in medical sciences. Nevertheless, the estimations utilizing this method are biased and underestimate the variability of the desirable statistic. The next method is the regression imputation which takes advantage of other variables to impute the missing values [8,10,11]. Particularly missing values are imputed by the predictions of the regression line. This specific regression line has as an outcome the variable with missing values and as predictor variables have the other variables with complete data. The purpose is the estimation of regression lines and the prediction of missing values using the values of predictor variables. The resulting predicted dataset is complete data. This approach ensemble the information of the other variables with the missing variable to predict the missing outcome. This method produced unbiased estimates of the mean when the missingness is MCAR or MAR. Nevertheless, the data structure is not preserved which means that the correlation of the variables and their variability is biased. Next, the k nearest neighbor imputation method is a statistical learning algorithm that imputes the missing values using the k nearest observations[12]. It has risen in popularity due to the ascending prominence of machine learning models[12,13]. The estimation of the number of nearest neighbors is conducted through cross-validation and then after the optimal selection of k, the missing values are predicted. Particularly it finds the k nearest observation as a specific missing observation and then using the k observation it predicts the missing values. This method is nonparametric in opposition to the regression, and it uses the relation of the missing variable with the other. It suffers from ordinary machine learning problems like overfitting. Also, there is difficulty in choosing the best metric for finding the closest observations because Euclidean metric is not appropriate to factor variables but there are a lot of solutions to this problem like dummy variables and byte encoding, and more. More methods like this kind can be employed like random forests, artificial neural network ensemble

boosting techniques like extreme gradient boosting and more leading to the conclusion that a missing value problem is a prediction problem[13]. Also, those machine learning techniques can infused with other statistical methods like multiple imputation for better results. Stochastic regression is a method that merges the regression imputation method with a random error [8,10,11].In other words, it unites the deterministic part of the prediction with randomness resulting in the preservation of the variability and the data structure of the initial dataset. Practically after predicting a missing value a random error is added to the result. This method has better results than the other because it keeps the data structure and the variability of the data. This error term follows normal distribution with zero means and a variance of the residuals leading to the steadiness of the variability. This stochastic regression produces unbiased parameter estimations under the assumption of MCAR and MAR. Nevertheless, a disadvantage of the parametric regression method is that some of the assumptions are frequently violated. For example, the errors must follow a normal distribution with invariable standard deviation, but those assumptions are not always valid. Another disadvantage this method has relative to the upcoming prominent and sophisticated methods is that the standard errors are not being adjusted from this method[8,10,11]. A standard error adjusting is necessary because the predicted scores are just forecasts for the true missing values leading to reduced and unsuitable standard errors which influence the statistical tests. The upcoming methods like multiple imputation and likelihood estimation remedy the previous problem.

2.2.4 Multiple Imputation and likelihood methods

Those two general methods are considered sophisticated and modern approaches to the missing values problems[14,8,10,11]. They can solve MCAR and MAR problems utilizing distinct, cumbersome, and effective algorithms and methods. The results that they produced are valid. In other words, the parameters estimated are unbiased under the assumption of MCAR and MAR missingness. Also, they solve the problem related to the variability and the preservation of data structures and they adjust the issue with the standard deviations by not treating a specific prediction as its true value. Finally, these two algorithms are like two families of various methods. For example, algorithms for multiple imputations are numerous two of them are data augmentation and multiple imputations by chained equation, the general concept of the maximum likelihood method is to estimate the so-called full information maximum likelihood one way to accomplish this is to utilize the expectation-maximization algorithm[8,10,11]. Nevertheless, those methods produce biased results when the missingness falls under the category of MNAR data, but the bias is less than the bias of the single imputation methods.

Multiple imputation is a method that in the end produces different independent datasets[8,10,11]. Those imputed final datasets are complete, and they differ only in the imputed missing values. Then after the generation of various datasets, the estimated parameters of each analyzed dataset are pooled in a specific manner. Particularly there are three steps for multiple imputation (MI): imputing the data, analyzing the data, and pooling the results. The distinct aspect

of various multiple imputation methods at the most times is the way of imputing the data. In the imputation phase, several copies of datasets are produced (a proposed number is 20) and each of the datasets has different estimates of the missing values. The most used method of estimating the missing values is data augmentation for normally distributed data[8,10,11]. There are two steps for this algorithm the first is the imputation step (I–step) and the second is the posterior step (P-step). To begin with, the I-step is like the stochastic regression imputation. For every incomplete variable a regression is constructed with the outcome of the variable with missing values, every regression has its estimated coefficients and its covariance. Then using this regression with the addition of a random error same as the random error of stochastic regression, the imputed values are generated. In other words, the imputed values are derived from the predicted values adding a random error to introduce variability to the model. Then after the imputation of the dataset, a complete dataset is proceeded to the P-step which through Bayesian algorithms each estimated coefficient with its covariance matrix is updated using the whole filled dataset. After the generation of those means and covariance matrixes, a random error related to the dimension is added to the estimates. This process creates a different regression than used in the first step. Then this regression is used to impute the previously imputed missing values with the addition of a random error. So, this process goes on. To summarize the whole algorithm an analyst can break down this algorithm into two parts, the first part is the imputation, and the next part is the updating of imputation algorithms. In other words, first, the imputation is done through stochastic regression and then the first complete dataset is used to recalculate this stochastic regression through Bayesian principles. After that, a second complete dataset is created and that goes on. After the construction of many datasets which have unique but not independent estimates this procedure finally can stop. Another important issue is that the datasets that will eventually be chosen to be analyzed need to be independent. This is practically achieved via repeating the algorithm numerous times and choosing one data set every for example 100 or 200 datasets. In other words, someone can take the 200th dataset then the 400th and that goes one. The period that is needed to choose the dataset is estimated through the correlation of the estimated missing values. After that, the analysis phases are conducted with the same methods and then the estimates are summarized or pooled in one. So, a numeric difference in each parameter estimation is introduced which adjusts the problem of standard deviation in a single stochastic imputation. Therefore, after the completion of each analysis, several desirable estimated parameters have been generated. The pooled estimated parameter is created by calculating the mean of those estimated parameters [8,10,11]. The hard side of the pooled step is the estimation of the standard deviation of this pooled parameter. To calculate the desirable standard deviation an analyst has to calculate the within and between variances. The within and between variances are described as

$$W = \frac{\sum_{i=1}^{m} SE_i^2}{m} \qquad \text{and} \qquad B = \sum_{i=1}^{m} \frac{(\theta_i - \bar{\theta})^2}{m-1}$$

Where SE_i^2 is the variance of each parameter estimation for $i = 1,2, \dots, m$ imputed datasets, θ_i is the estimated parameter and $\bar{\theta}$ is the pooled estimated parameter

which mathematically is described as $\bar{\theta} = \sum_{i=1}^{m} \frac{\theta_i}{m}$. So, the pooled standard error is function of the previous quantities and is given as

$$SE = \sqrt{W + B + B/m}$$

The single imputation method underestimates the standard error because it assumes that the predicted value is the real one but in the multiple imputation, this aspect is solved with the between variance. There are many versions of the previous, some of them assume for example that there is only one-factor variable with missing values and instead of regression a generalized regression model is used and more … Multiple Imputation by Chained Equations (MICE) [14] is a multiple imputation which follows the same phases of the data augmentation analysis. Specifically, it has 3 phases the imputation phase, then the analysis phase, and finally the pooling phase. The last two phases are the same as the previous algorithm and in general, as it is said above those 3 phases are the same for every multiple imputation algorithm. Therefore, the imputation algorithm is different. This algorithm is the most acceptable, understandable, and simplest algorithm under the powerful and effective family of multiple imputation and likelihood methods their results are unbiased for MCAR and MNAR. This method is easily used and some packages conduct this type of multiple imputation. So, the algorithm is, first impute all the missing values with their variable means or the most probable value for factor variables. Then starting from the first variable with missing values, build a regression that has as covariates the rest of the variables and replace the imputed values with missing values. After the application of the regression, predict the missing values. Then proceed to the next variable with missing values, replace all the imputed values with missing and build a regression using the other variable (with imputed values) as covariates then predict the outcome and replace the missing value of the dependent variables. After the employment of this concept, the rest of the variables start again in a second cycle same as the first. Particularly, again replace the imputed values of the first variable with missing values then employ a regression with covariates the other complete variables. After the prediction of the missing values of the first variable, keep doing this for the rest variables and then a second cycle is finished. Generally, 10 cycles are conducted but an analyst can define another number of conducted cycles which is related to the successive distance cycle datasets[14]. For instance, if the distance of data frame of cycle x and the data frame of cycle $x + 1$ is lower from a certain value then break the loop. Finally, the final cycle is the desirable complete dataset. In practice several datasets are needed, to create those datasets this procedure starts from the beginning. Nevertheless, the first imputation with the means can be a stochastic imputation or the regression imputation can be a stochastic one to change the final results, or someone can change the order of the variable predictions, or an analyst can employ all 3 ways together. Again, there are a lot of ways in which MICE can be applied. Some disadvantages of this method are that again the assumption of regression might not hold and the calculations that are needed for 10 cycles with numerous variables are enormous. Finally, an analyst can use MICE to employ it like a single imputation method, for example, they can create just the first complete data set and take only that. Another powerful change that can be implemented is the replacement of the regression

model with a general machine learning model like random forests or artificial neural networks and more.

Likelihood methods are trying to estimate the full information likelihood [8,10,11], as it is called in literature, which is the same normal likelihood using all available data. The estimation process of the likelihood is the same either in complete or incomplete data. The maximum likelihood process of incomplete data is to calculate the estimators of the desirable parameters that maximize the likelihood or respectively the log-likelihood. In the non-complete dataset framework again the calculation of the parameter's estimators just requires the available dataset ignoring the missing values. This method does not produce a complete dataset like multiple imputation, but it just calculates the preferred likelihood estimators using all the available datasets. The problem which arises is the computation of standard errors which are acquired using the expected or the observed information matrix. The expected information matrix uses only the complete dataset, but the observed information matrix is dependent on the missing values. The expected information produces standard errors which require the MCAR missingness and the observed information matrix produce standard error which are required for both MCAR and MAR assumption. The observed information matrix is the desirable matrix and the option for the observed information matrix is available in various software packages. This information matrix is derived from the one minus Hessian matrix and the standard errors of variables are the root of the diagonal inverse information matrix. In the observed information matrix, the sample statistics are used instead of the expected sample statistics. As the multiple imputation, the likelihood method for missing values requires an algorithm to optimize the likelihood even for a simple parameter estimation. One algorithm that is used for the optimization is the expectation-maximization algorithm (EM). Practically, the mean of using only the available data ignoring the missing values is that every individual likelihood uses only the available observed variables. It is something like pairwise deletion for every observation. Mathematically, let's assume that data follow a multivariate normal distribution, and the purpose is to compute the mean and the covariance matrix. In the complete case scenario, the log-likelihood which is requested to be maximized is
$\log L = \sum_{i=1}^{n} \log L_i = \sum_{i=1}^{n} \log f_i(x_1, x_2, \dots, x_p/\mu, \Sigma)$ where n is the number of observations, $x_1, x_2, \dots, x_p$ are the observed values of the variables $X_1, X_2, \dots, X_p$, μ $(p \times 1)$ is the mean of the variables and Σ is the covariance matrix $p \times p$. Now let assume that an observation has missing values in one variable. Specifically, without the loss of generality, the kth individual has missing value on the variable X_j where $j = 1, 2, \dots, p$, their specific log likelihood is

$\log L_k = \log f_k(x_{1,} x_2, .., x_{j-1}, x_{j+1,} \dots, x_p/\mu^j, \Sigma^j)$ where μ^j is the expected mean and Σ^j is the covariance matrix of $X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_p$. In other words, μ^j is the complete μ without the jth term and Σ^j is the complete Σ without the jth row and column. Thus, the whole likelihood under the specific missingness becomes $\log L = \sum_{i=1}^{n} \log L_i = \sum_{i=1, i \neq k}^{n} \log L_i + \log L_k$ where

$\sum_{i=1, i \neq j}^{n} \log L_i = \sum_{i=1, i \neq k}^{n} \log f_i(x_1, x_2, \dots, x_p/\mu, \Sigma)$ and

$$\log L_k = \log f_k(x_1, x_2, .., x_{j-1}, x_{j+1}, \dots, x_p / \mu^k, \Sigma^k).$$

In the general missing value scenario, every observation has its log-likelihood like the kth observation. In practice, optimization algorithms are used even for relatively small missing value problems or even scenarios in which a small number of variables have missing values. The expectation–maximization algorithm (EM) is an algorithm that is used to optimize the likelihood of missing values. The EM is used in every type of parameter estimation like mean, covariance, coefficient estimation, and more. This algorithm consists of two steps the E-step (expectation step) and M-step(maximization step) which are conducted iteratively. The iterative procedure begins with a pairwise deleted estimate of the mean and covariance matrix then using this initial estimation the E-step creates a group of regression models relative to the initial mean and covariance matrix estimation to predict the missing values. The E-step fills in the incomplete data using stochastic regression modeling. The M-step employs the ordinary maximum likelihood estimation to estimate the mean and the covariance from the fill-in data. Then again, the E-step recreates the stochastic regression models utilizing the latter estimation of mean and the covariance matrix. Also, in the E-step the missing values are refilled with the predictions from the stochastic regression and that goes on until there is a convergence in the mean and covariance matrix. The convergence of the mean and covariance matrix is achieved when there is no substantial difference in both estimations between two sequential M-steps. Conclusively in all M-step, the maximum likelihood estimators are calculated using the E-step filled data resulting in the final maximum likelihood which is eventually the expected one. This sequence of maximum likelihood estimators converges with the true maximum likelihood estimator. Nevertheless, the EM is not an algorithm for data imputation but for the calculation of the maximum likelihood estimator. This final maximum likelihood estimator is unbiased of the maximum estimator under the assumption of MCAR and MAR.

Both algorithms multiple imputation and likelihood estimation are most frequently used and they are unbiased for the MCAR and MAR but MNAR missingness[8,10,11]. Nevertheless, their estimations in the MNAR framework are less biased than the single imputation estimations. Finally, both methods produce consistent, unbiased, and canonical estimators. Apart from those two methods, there is another prominent general approach which is called inverse probability weighting. In this approach, there are several inverse probabilities weighting methods that can efficiently be applied to various missing situations.

2.3 The competing risks with missing cause of failure model

Competing risks with the missing cause of failure is a frequent phenomenon that arises when there are several event types and some of them are missing. Specifically, only the event type variable has missing values and not the predictor variables. Several algorithms rectify the missing cause of failure issue under the missing at-random assumption in the semi-parametric or Bayesian-parametric framework[9]. Some simple and easily conducted algorithms are the complete case method or the creation of another new event from the missing event but those

methods sometimes are biased and misleading or in the best scenario the complete case method just reduces the power. Consequently, better approaches are essential to have valid and trustworthy results. Initially, a multiple imputation approach to handle the missing cause of failure has been mentioned by Lu and Tsiatis [9,15] which imputes the missing cause of failure with a probabilistic competing event model. A second strategy is to use the concept of weighting in the likelihood. Specifically a maximum pseudo partial likelihood is employed via the assignment of weights to the observation, the weights are a function of the same probabilistic competing event method of Lu and Tsiatis [15,9], this method is mentioned by Bakoyiannis et al[16]. The third approach is to utilize a data augmentation approach using Bayesian methods to remedy this problem [10,17,18]. Some proposed methods are based on the EM algorithm (Craiu and Duchesne 2004)[19], the partial-likelihood approach (Goetghebeur and Ryan 1995) [20], and the augmented inverse probability weighting (Hyun et al 2012)[21].In this master thesis, the first three approaches are efficiently described.

The first method represented is multiple imputation which predicts the missing outcome with a stochastic regression for every dataset[9,15]. Let assume that T donates the failure time and C the event type or cause of failure. According to the competing risk approach there are several event types that it is specifically assumed that there are just two types of events. Eventually C is equal to 1 when the first cause of failure has happened or is equal to 2 for the second event occurrence. For each individual i $(i = 1,\ 2, \dots, n)$ there is T_{i1}, T_{i2} which are related to two latent time failures and the censoring time U_i. The censoring time is independent from the two event times, and it falls under the category of the independent right censoring. Consequently, for each subject i is observed (T_i, C_i, Z_i) where $T_i = \min(T_{i1}, T_{i2}, U_i)$, $C_i = 1,2$ and 0 when the observation is censored and Z_i the other variables. The cause-specific hazard and the cumulative incidence function is described as

$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t \le T \le t+\Delta t, C=k | T \ge t)}{\Delta t}$ and $F_k(t) = P(T \le t, C = k)$ further details are given in the first sub paragraph. Let assume R a missingness indicator which is equal to 1 if the C is observed and 0 if it is not. If $R = 0$ it is assumed that $C > 0$ and the type of missingness is missing at random (MAR) meaning that the probability of missingness is independent of the true missing cause of failure but only to observed variables. Also, some auxiliaries covariates X_i are used because they may are related to the missingness probability and not to the hazard function. Mathematically the probability of missingness for the ith individual is given as : $P(R_i = 0 | C_i, C_i > 0, W_i) = P(R_i = 0 \mid C_i > 0, W_i)$ where $W_i = (T_i, Z_i, X_i)$.

Multiple imputation consists of 3 phases imputation, analyzing and pooling. In the imputation phase several m datasets are created which they differ only in the imputed missing cause of failure. The missing cause of failure for the first event is imputed from a Bernoulli distribution with probability

$P(C_i = 1 \mid R_i = 0,\ C_i > 0, W_i)$. Under the MAR assumption the probability $P(C_i = 1 \mid R_i = 0,\ C_i > 0, W_i)$ is equal to $P(C_i = 1 \mid W_i) = \pi_1(W_i)$. It is proposed that a logistic regression is used in order to model the $\pi_1(W_i)$, using as outcome

the variable Y_i which is equal to 1 if $C_i = 1$ and 0 if $C_i = 2$ (considering the first event as event of interest) and as covariate all possible and appropriate combination of W_i. Furthermore, it is suggested that the introduction of high order polynomial, splines, fractional polynomial, interactions, and other variable modeling strategies are required in order to accumulate an effective and valid predictive model. Therefore, after the modelling, the vector $g(W_i)$ of components W_i and its combination of all plausible variable modelling $f(W_i)$ is used with aim of acquiring the desirable probabilities. Particularly, utilizing the logistic model it is known that $logit\big(\pi_1(W_i)\big) = g(W_i)'\theta$ where θ is the vector coefficient. Eventually, $\pi_1(W_i) = \frac{\exp(g(W_i)'\theta)}{1+\exp(g(W_i)'\theta)}$. Consequently, based on the probability of $\pi_1(W_i)$ the m datasets are imputed. Firstly, run the logistic model with only the complete dataset in order to acquire $\hat{\theta}$ and $\widehat{Var(\theta)}$ where $\widehat{Var(\theta)}$ is the inverse information matrix. For the dataset imputation, stimulate θ^* from $N(\hat{\theta}, \widehat{Var(\theta)})$ which is the posterior probability under the Bayesian framework. Then calculate $\pi_1(W_i) = \frac{\exp(g(W_i)'\theta^*)}{1+\exp(g(W_i)'\theta^*)}$ and then assign to the missing cause of failure the first event with probability $\pi_1(W_i)$ and the second event with probability $1 - \pi_1(W_i)$. This dataset imputation is conducted m times after that the analysis phase is carried out which is the typical analysis because each dataset is complete. In each analysis the same models or algorithms are applied resulting in several m estimate effects. The pooling phase is conducted via the same formula as the multiple imputation algorithm. Without the loss of generality given that the estimated effect of the ith analysis is θ_i for the first cause of failure, the pooling average is $\bar{\theta} = \sum_{i=1}^{m} \frac{\theta_i}{m}$ and for its standard error the between and within variance are needed. The between variance is a metric for deviation of θ_i from $\bar{\theta}$ and it is described as $B = \sum_{i=1}^{m} \frac{(\theta_i - \bar{\theta})^2}{m-1}$. The within variance is the mean of standard errors of each estimate effect that is $W = \frac{\sum_{i=1}^{m} SE_i^2}{m}$. Therefore, the standard error of $\bar{\theta}$ is the $SE = \sqrt{W + B + B/m}$. This multiple imputation algorithm for missing cause of failure is very flexible because eventually the analysis is conducted in complete datasets and it is unbiased under MAR assumption. The disadvantage is that it is very numerically heavy because sometimes more imputed datasets than average are essentially needed and sometimes there are more causes of failure than 2. Finally, this method is also applicable to semi-parametric subdistribution hazards and in the parametric survival framework.

The second approach is the so-called maximum pseudo-partial-likelihood estimation method (MPPLE) which utilizes the cause-specific semi-parametric regression [16]. This method assigns weights to the observations with a missing cause of failure and then the maximization of the likelihood is conducted normally. Specifically, let's assume that the first cause of failure is the event of interest, as it was mentioned in the cause-specific hazard paragraph that several analyses are conducted by treating the other events but the event of interest as censor events in each analysis. When the outcome is missing, those observations with the missing event are duplicated, in the first duplication the events of that observation are replaced with the event of interest, and in the second copy, the events are

treated as censored. In other words, the observations with missing events are doubled and in the first batch they are imputed with the outcome of interest, and in the next, they are treated as censored. The weighting of the outcome of interest imputed observation is the probability of the desirable event occurrence and the weight in the same subject which is treated as censored is the one minus the probability of the event of interest occurrence. This probability is the same as the one used for the treatment of the missing cause of failure in multiple imputations. Then after the event or censor imputation and the assignment of the specific probabilities, the maximum likelihood is ordinary maximized. The normal Cox regression is applied resulting in the acquisition of the coefficient estimation. Nevertheless, one minor problem is that the standard errors of Cox regression are invalid which leads to the utilization of bootstrap methodology to acquire the standard errors.

The third approach is the Bayesian data augmentation method which comes in very handy with the requested model of this thesis[10,11,17]. Inherently data augmentation is a type of multiple imputation algorithm but when the coefficients of the requested parametric model are estimated using Bayesian statistics then the data augmentation can be included in the Metropolis Hasting algorithm. Consequently, there is no need to extract one every turn one dataset because the estimation of the coefficients is conducted after the creation phase of the dataset in the Metropolis Hasting loop. Specifically, as was mentioned, in the first cycle the values of missing causes of failure are simulated from their posterior predictive probability with the help of the logistic model and then using the observed and the imputed observation the parameters of the logistic regression are updated using the parameter's posterior probability. After that a complete first cycle dataset has been created, then using this complete dataset and Bayesian statistics the first values of the requested coefficients are calculated. Then the same methodology is repeated for numerous cycles and after some burn-in period, the analyst takes the rest values of the coefficients. In the end, the coefficients of each cycle are aggregated resulting in the final estimations. Using data augmentation and Bayesian parametric modeling, there is no need to keep in software memory every few turns the complete dataset but only the coefficients.

In conclusion of the 3 methods rectifying the missing cause of failure, only the last two are used in this thesis and further information about them is given in the next chapters. All approaches use a logistic regression model to remedy the missing cause of failure issue but there is a possibility that this model might not be a great fit for the data resulting in bias introduction, thus it is essential to try fitting the best possible logistic model. The advantage of the last two approaches is that they do not store numerous complete datasets as the first one. Nevertheless, the first approach is far more flexible than the others because the second approach is valid for cause-specific semi-parametric regression and in the third approach sometimes it is hard to calculate the posterior distribution of the missing values, parameters of the logistic regression, and the requested coefficients. All methods that are described are at least valid for independent right censoring and under MAR assumption.

3 Bayesian parametric survival analysis

3.1 Bayesian Analysis

3.1.1 Bayesian Inference

Bayesian analysis is about assigning a probability model to a set of data and making inferences from a probability distribution on the parameters of the model or unobserved quantities such as predictions [18,22,23]. Bayesian inference is based more on probability theory rather than the frequentist's inference because the parameters in Bayesian analysis are considered random variables and they follow a certain distribution. The whole concept of the Bayesian inference is first to assign a probabilistic model to the parameter and then update it using new information which practically this new information is the data. In other words, the main purpose of Bayesian analysis is to refresh the distribution of the desirable parameters with valuable information. Then an informed distribution of the parameters is extracted, and this updated distribution is the main purpose of the analysis. This so-called posterior distribution of the wanted parameters and contains all the inference [18,22,23]. In other words, after calculating the posterior distribution the inference is made by summarizing the updated distribution. The advantages of Bayesian statistics are the more probabilistic approach because in the end it is known exactly the parameter probability to be included in a specific interval. Another advantage is that there is no problem with data sufficiency because the exact updated probability of the parameters is calculated utilizing general probability theorems and not by estimations and asymptotic theorems as the frequentist's approach. The main disadvantage is that Bayesian inference is numerically heavy, and the process of inference requests some assumptions and most of the time researchers are not even confident that those assumptions are valid.

The steps for the conduction of Bayesian inference are firstly assign a probability to all observed and unobserved quantities. The probability should be related to the prior knowledge or the nature of the data. Then the calculation of the desirable posterior distribution of parameters conditioning on the data (information) is conducted to summarize the results. Finally, the evaluation of the model fit is carried out to check the goodness of fit [18,22,23]. The calculation of the posterior distribution by hand is almost always impossible or sometimes it needs cumbersome calculations. Nevertheless, with the help of simulation methods like inverse probability method, accept reject, Metropolis Hasting, and more the posterior distribution is efficiently summarized. Therefore, the inference on the parameter θ or on the unobserved quantities like the prediction of certain values are made from a probabilistic approach. Specifically, the probability model is updated from the observations y resulting to the a posterior distribution $p(\theta|y)$ which it's random variable θ given y and to the posterior predictive distribution $p(\tilde{y}|y)$ which is the distribution of the predicted values conditioning on the observed values.

The Bayesian statistics started from the formula of conditional probability and the connection with the joint. Specifically, the joist distribution of θ and y is

equal to $p(\theta, y) = p(\theta|y)p(y)$ and to $p(\theta, y) = p(y|\theta)p(\theta)$ and those two relations leads to the bayes rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Where the $p(y|\theta)$ is the probability of the y realization given θ which is basically the so called likelihood, $p(\theta)$ is the prior probability of the parameter θ which is virtually is the initial belief about the parameter and $p(y)$ is the probability of the data which mathematically is equal to $\sum_\theta p(y|\theta)p(\theta)$ or $\int_\theta p(y|\theta)p(\theta)d\theta$ in case of θ being discreet variable or θ being continuous. The formula $p(\theta, y) = p(y|\theta)p(y)$ is used and then the sum or the integral is for all possible values of θ is calculated. The $p(y)$ is named as normalizing constant because it is an invariable quantity which the division of this number from the $p(y|\theta)p(\theta)$ results in the posterior distribution and without that, the $p(y|\theta)p(\theta)$ is not a distribution [18,22,23]. In the most times $p(y)$ is hard or even impossible to calculate; therefore, is omitted because it is not function of θ and then the posterior distribution is analogous of the likelihood and the prior distribution which mathematically is

$$p(\theta|y) \propto p(y|\theta)p(\theta) = h(\theta)$$

The posterior predictive distribution $p(\tilde{y}|y)$ is frequently used for predictive problems or for treating missing values problem[18,22,23]. The $p(\tilde{y}|y)$ is called posterior predictive distribution because is conditioned to the observed values y .The posterior predicted distribution derives from the upcoming formula

$$p(\tilde{y}|y) = \int p(\tilde{y}\,\theta|y)\,d\theta = \int p(\tilde{y}|\theta, y)p(\theta|y)\,d\theta = \int p(\tilde{y}|\theta)p(\theta|y)\,d\theta$$

Which indicates that the probability of the predicted values given the parameter θ is independent of the observations y. The posterior probability distribution $p(\theta|y)$ is related to the data through $p(y|\theta)$ which is so called likelihood function and that is the reason the Bayesian analysis is related to likelihood properties. The ratio of $p(\theta|y)$ between two points θ_1 and $\theta_2$is called posterior odds of θ_1 compared to θ_2 [18,22,23]. Mathematically the posterior odds of θ_1 and θ_2 is

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{\frac{p(y|\theta_1)p(\theta_1)}{p(y)}}{\frac{p(y|\theta_2)p(\theta_2)}{p(y)}} = \frac{p(y|\theta_1)p(\theta_1)}{p(y|\theta_2)p(\theta_2)} = \frac{p(y|\theta_1)}{p(y|\theta_2)}\frac{p(\theta_1)}{p(\theta_2)}$$

Which $\frac{p(y|\theta_1)}{p(y|\theta_2)}$ is called the likelihood ratio and $\frac{p(\theta_1)}{p(\theta_2)}$ is the prior odds of θ_1 and θ_2. The concept of posterior odds is usually applied to a discreet parameter θ and sometimes if θ is continuous the posterior odds of the θ_1 interval versus θ_2 interval is possibly requested one.

The inference about the parameter θ is made by the posterior probability distribution. Therefore, all the possible information from the posterior probability distribution is essential to be extracted in order to conclude some inferences. To begin with, a histogram or a density plot of $p(\theta|y)$ contains

valuable graphical information about the updated parameter θ [18,22,23]. Those plots illustrate the symmetry or the skewness of distribution, and also, they demonstrate the mode or modes of the posterior probability distribution if it is unimodal or not. Nevertheless, measures of centrality and variability are needed such as mean, median, mode, standard deviation, quantiles, and others. In most cases, the posterior probability distribution is not a closed form or if it is, it is so complicated that those measures of centrality and uncertainty need cumbersome calculations. Therefore, this problem is solved from the simulation theory and those central and variability metrics are easy to calculate through the employment of simulation methods. This simulation concept provides this nice to have flexibility for metric calculations even when complex transformations are previously done. In most cases, a simulation method is conducted and metrics like mean, standard deviation, and quantiles are represented. An important issue is measuring the posterior uncertainty, this is conducted through the calculation of posterior intervals [18,22,23]. Specifically, the usual strategy is to present quantiles of $p(\theta|y)$, technically a range is given which is called central interval. The posterior probability of central interval is equivalent to $1-a$ and contains all the values which the minimum value has posterior probability lower than the $a/2$ and the maximum value has the corresponding probability bigger than $1-\frac{a}{2}$. Practically, let assume (c_1, c_2) be the central posterior interval of $100(1-a)\%$, in terms of posterior interval the c_1 and c_2 are found from the $\int_{c_1}^{c_2} p(\theta|y)d\theta = 1-a$ but virtually the c_1 and c_2 are calculated using the cumulative posterior probability (or its corresponding integral). The formula for the range of the central posterior interval is $\int_{-\infty}^{c_1} p(\theta|y)d\theta = \frac{a}{2}$ and $\int_{c_2}^{+\infty} p(\theta|y)d\theta = 1-\frac{a}{2}$. For some specific posterior distributions such as normal or binomial the central posterior interval is conveniently given by the corresponding tables. Nevertheless, as it is mentioned in most cases the posterior distribution is too complex to have a specific table for quantiles, and in those cases, those quantiles are extracted from the simulation sample like the ordinary estimated sample quantiles. In addition to the central posterior interval, another interval sometimes is given which is called as highest posterior density region. This so-called highest posterior density region is the typical posterior interval but with the addition that the values inside the interval have higher density than the values outside the interval. In other words, the bounds of the interval or intervals have a density bigger than a fixed number, and the interval or just a union of them has posterior probability of $1-a$. This highest posterior density interval is preferred to the central posterior interval when the distribution is not unimodal or is highly skewed. Nevertheless, those two regions are the same when the distribution is symmetric with zero skews [18,22,23].

The selection of prior distribution is very important in Bayesian statistics because prior distributions sometimes are very influential to the posterior distributions. The prior distribution is the prior or initial knowledge of the parameter before the observation of new data-information. So basically, sometimes the researchers have a specific initial knowledge about the parameter and sometimes they do not. When there is specific prior information about the parameter the prior distribution has lower variance; as a result, bigger information and when there is not specific information the prior distribution is

flat with high variance leading to low information [18,22,23]. The prior distribution in the first situation is more informative and influential than the prior distribution in the second situation. The posterior distribution is more dependent on the data when the prior distribution is noninformative and less dependent on when it is informative. In other words, the posterior distribution is similar to the corresponding estimated frequentist distribution when the prior is noninformative and more distinct when it is informative. Nevertheless, the prior distribution should include all the plausible values of θ either it is informative or not. The prior distributions also can break down to two categories proper and improper ones. Improper prior distributions are virtually analogous to a fixed value c for $\theta \in (-\infty, +\infty)$ or mathematically it is described as $p(\theta) \propto c,\ \theta \in R$. As a result, this prior is not a distribution because it's sum or integral for values of θ is not 1 and the posterior distribution is analogous only to likelihood function. Those improper prior distributions are part of noninformative family. On other hand proper distribution are those whose sum or integral for viable θ is equal to 1. Another popular type of priors is Jeffrey prior which is based on the invariant principle. This type of prior used to create noninformative prior related to one-to-one transformation φ=h(θ) $\underset{1-1}{\Longleftrightarrow} h^{-1}(\varphi) = \theta$. By the application of chain rule the prior distribution of φ is $p_{\varphi}(\varphi) = p_{\theta}\left(h^{-1}(\varphi)\right)\left|\frac{dh^{-1}(\varphi)}{d\varphi}\right|$. The general concept of Jeffrey's principle is that there is similar corresponding rule which is applied to the prior of θ and to the prior φ. The Jeffrey's noninformative prior is defined as $p(\theta) \propto \sqrt{J(\theta)}$ where $J(\theta)$ is the Fischer information matrix which is described as :

$$J(\theta) = E\left(\left[\frac{d\log p(y|\theta)}{d\theta}\right]^2\right) = -E(\frac{d^2\log p(y|\theta)}{d\theta^2})$$

So in order to find the $J(\varphi)$, the $h^{-1}(\varphi) = \theta$ relation is used leading to the invariant parametrization principle. Particularly,

$$\sqrt{J(\varphi)} = \sqrt{J_{\theta}\left(\, h^{-1}(\varphi)\right)}\ \left|\frac{dh^{-1}(\varphi)}{d\varphi}\right|$$

Finally, a prominent type of priors is those which called conjugate priors[18,22,23]. Basically, the formal definition is that the P is conjugate class for F if $p(\theta|y) \in P$ for all $p(\cdot\,|\theta) \in F$ and $p(\cdot) \in P$ which F is sample distributions of likelihood and P is sample distribution of prior. Basically, it says that the prior and the posterior belongs to same family under the conjugate property. Specifically, the naturally conjugacy is when the P and F is the same family. This natural conjugacy property is applied for the distributions which belong to the exponential family. Virtually if the prior and the likelihood distribution belong to the exponential family then the posterior belongs to exponential family, and it has a closed form. The proof is : Let $y = (y_1, y_2, \ldots, y_n)$ be identical and independent observations which are belong to an exponential family and θ be a multidimensional vector. Therefore,

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\varphi(\theta)^T u(y_i)}$$

Where $\varphi(\theta)$ and $u(y_i)$ are same the dimension as θ. The vector $\varphi(\theta)$ is called natural parameter of the exponential family F. The likelihood of the vector y is

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} f(y_i) g(\theta) e^{\varphi(\theta)^T u(y_i)}$$

$$= g(\theta)^n e^{\sum_{i=1}^{n} \varphi(\theta)^T u(y_i)} \prod_{i=1}^{n} f(y_i)$$

which basically results to

$$p(y|\theta) \propto g(\theta)^n e^{\varphi(\theta)^T \sum_{i=1}^{n} u(y_i)}$$

The $t(y) = \sum_{i=1}^{n} u(y_i)$ is sufficient statistic for the θ which means that $p(y|\theta)$ is depended through the observations only from their sum. Now let the prior belong to exponential family.

$$p(\theta) \propto g(\theta)^{\eta} e^{\varphi(\theta)^T \nu}$$

Eventually, the posterior distribution is written as

$$p(\theta|y) \propto g(\theta)^{\eta+n} e^{\varphi(\theta)^T (\nu + t(y))}$$

This leads to the desirable result that the posterior belongs to the exponential family. Conclusively, this natural conjugacy closed-form property is used in the univariate analysis, but the general conjugacy property is used to determine the form of the conditional posterior distributions in the multivariate Bayesian analysis. Typically, the selection of prior is conducted with the help of a bibliography and the general community suggestions[18,22,23]. Finally, by and large, an analyst starts from an informative before a noninformative and vice versa to find the optimal parameters for the algorithm convergence or they conduct various sensitivity analyses to observe the impact of certain priors.

Practically in statistical problems, more than one parameter or unobserved quantities are involved. Therefore, Bayesian analysis has many advantages over other inference methods in terms of the multiparameter problems. Conclusions for this multiparameter problem are drawn only for one or some parameters at a time. The basic aim of those types of problems is to acquire the marginal posterior probability of the parameters of interest which is the posterior probability of the desirable parameter. Theoretically, to acquire the marginal posterior probability first the joint posterior probability is defined then the integral of this joint distribution is calculated[18,22,23]. Or if simulations are the key to this issue, an analyst draws simulation from the joint probability then they make inferences only from the desirable parameter simulations, ignoring the other simulations. Those parameters that are indifferent in terms of inference which are called nuisance parameters are necessary to create a valid model and acquire the marginal posterior distribution of the parameters of interest. Practically the joint posterior density of two parameters θ_1, θ_2 is given from

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2) p(\theta_1, \theta_2)$$

Let assume that θ_1 is the parameter of interest and θ_2 is the nuisance parameters. As it is previously mentioned the inference is made by the marginal posterior distribution of θ_1 and the marginal distribution of θ_1 is derived from

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)\, d\theta_2$$

One another way to write $p(\theta_1|y)$ is

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2|y)\, d\theta_2$$

Which shows that $p(\theta_1|y)$ is related to the posterior distribution of nuisance parameter and the conditional posterior distribution of θ_1 given the nuisance parameter. Alternatively, $p(\theta_1|y) = E_{\theta_2|y}[p(\theta_1|\theta_2, y)]$ which is the average of the condition posterior distribution given the nuisance parameters multiplied by the weighting function the marginal posterior distribution of θ_2 over the θ_2. In the most cases, the integral is basely not calculated but an alternatively approach is conducted which first θ_2 is simulated from its posterior marginal distribution and then the θ_1 is simulated from its conditional posterior distribution given the simulated value θ_2 [18,22,23].

Conclusively, the Bayesian inference is conducted through the posterior distribution of θ which in the most cases is impossible or very to hard to be calculated leading to the implementation of simulation approaches. From the simulations, the inference is drawn via the centrality and variability measures of the sample. A basic approach to solve Bayesian problems is first to write the likelihood $p(y|\theta)$ as a function of θ. Then write the posterior density $p(\theta|y) \propto p(y|\theta)p(\theta)$ and try to determine a prior distribution either informative or noninformative, proper or improper or if it possible and viable a conjugate prior. The prior is written like the likelihood function, as function of θ. Then after having $p(\theta|y) \propto p(y|\theta)p(\theta) = h(\theta)$ either try to find the true form $p(\theta|y)$ from $h(\theta)$ and make inference from formulas or start simulations for $p(\theta|y)$ via $p(\theta|y)$ or $h(\theta)$. For the multiparameter problem the use of the condition posterior distribution is the key to solve the issue. Specifically, the parameters are simulated from the conditional posterior distribution given the observation or the other previously simulated (or initial values if it's the first simulation) parameters. In terms of the predictive posterior distribution, the first approach is to calculate theoretically $p(\tilde{y}|y)$ from the formula

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)\, d\theta$$

Or simulate $\tilde{y}$ from the $p(\tilde{y}|\theta)$ for a given θ which is simulated from the posterior probability distribution. So, for every θ, a predicted value $\tilde{y}$ is simulated.

One another important chapter in statistics either Bayesian or frequentist is hypothesis testing [22,23]. The typical concept of hypothesis testing is that there are two distinct decisions and one of them is the true one. Therefore, the analyst using the data must accept one of them. The ordinary idea is that the initial hypothesis is H_0 and the alternative is H_1. So, the two options are

$$H_0: \theta \in \Omega_o$$

$$H_1: \theta \in \Omega_1$$

The first one indicates that $\theta \in \Omega_0$ and the second one $\theta \in \Omega_1$ where Ω_0, Ω_1 are subset of the parametric space. In the easiest scenario Ω_0, Ω_1 are equal to $\{\theta_0\}$ and $\{\theta_1\}$ respectively. The solution to this problem is to calculate the posterior odds of θ_0 and θ_1 given the observations y $\frac{p(\theta_0|y)}{p(\theta_1|y)}$ [18,22,23] which is

$$\frac{p(\theta_0|y)}{p(\theta_1|y)} = \frac{\frac{p(y|\theta_0)p(\theta_0)}{p(y)}}{\frac{p(y|\theta_1)p(\theta_0)}{p(y)}} = \frac{p(y|\theta_0)p(\theta_0)}{p(y|\theta_1)p(\theta_1)} = \frac{p(y|\theta_0)}{p(y|\theta_1)}\frac{p(\theta_0)}{p(\theta_1)}$$

So, first a model for the likelihood is assumed in order for the likelihood ratio to be calculated and secondly two prior distributions for θ_0, θ_1 are presumed. After finding the $\frac{p(y|\theta_0)}{p(y|\theta_1)}\frac{p(\theta_0)}{p(\theta_1)}$, the interpretation of posterior odds is conducted. For example let assume that $\frac{p(\theta_0|y)}{p(\theta_1|y)} = c$ then the probability the of the parameter taking the value θ_0 is c times the probability of θ being θ_1. If $\frac{p(\theta_0|y)}{p(\theta_1|y)} > 1$ H_0 is chosen otherwise H_1. As it is observed, there is not p-values or first type error or significant level alpha and other [22,23]. The Bayesian interpretation is simpler and more logical that the frequentist approach because it is based directly on probabilities of a hypothesis being true given the data. Furthermore, an important difference is that the posterior odds is dependent on the prior information of θ and on the new information which is the sample. Another valuable metric of hypothesis testing is the Bayes factor BF of θ_0 versus θ_1 which in the previous example is

$$BF_{01} = \frac{p(y|\theta_0)}{p(y|\theta_1)}$$

In other word the Bayes factor is the likelihood ratio of two parameters, and it does not consider the prior distribution in simple hypothesis framework. If Bayes factor is incredibly large or small, then it can overcome the prior information. Another situation is that Ω_0, Ω_1 are intervals then after finding of the posterior distribution the posterior odds is calculated though the $p(\theta \in \Omega|y) = \int_{\Omega} p(\theta|y)d\theta$; namely

$$\frac{p(\theta \in \Omega_0|y)}{p(\theta \in \Omega_1|y)} = \frac{\int_{\Omega_0} p(\theta|y)d\theta}{\int_{\Omega_1} p(\theta|y)d\theta} = \frac{\int_{\Omega_0} p(y|\theta)p(\theta)d\theta}{\int_{\Omega_1} p(y|\theta)p(\theta)d\theta}$$

The interpretation is pretty much the same as the univariate example if $\frac{p(\theta \in \Omega_0|y)}{p(\theta \in \Omega_1|y)}$ is bigger than 1 then H_0 has better chance to be valid otherwise H_1 is more likely to be true. An introduction about the model's comparison is essential to be done, before continuing in the next basic form of hypothesis testing which is

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

There is more general concept apart from the hypothesis testing and that is the model comparison, particularly every hypothesis is a specific model [22,23]. So let assume that there are k models $M_1, M_2, \dots, M_k$ and the question is which model is most probable to be true. As a result, the posterior probability of each model given the data is estimated. The $p(M_i|y)$ for each $i = 1,2,3, \dots, k$ is estimated through the formula

$$p(M_i|y) = \frac{p(M_i)p(y|M_i)}{p(y)} = \frac{p(M_i)p(y|M_i)}{\sum_{i=1}^{k} p(M_i)p(y|M_i)}$$

Where $p(M_i)$ is the probability of model M_i is the true one and $p(y|M_i)$ is derived from $p(y|M_i) = \int_{\Omega_{\theta|M_i}} p(y|M_i, \theta)\, p(\theta|M_i) d\theta$. Basically $p(\theta|M_i)$ or $p(y|M_i, \theta)$ or both are usually different from the other $p(\theta|M_j)$, $p(y|M_j, \theta)$ for $i \neq j$. This concept of model comparison can expand to variable selection, checking the validation of various models, complex hypothesis testing and more. In this problem basically H_0 is M_0 and H_1 is M_1. Furthermore, the union of two sub spaces of θ is all the parametric space resulting in $p(H_0|y) = 1 - p(H_1|y)$. The Bayes factor is $BF_{01} = \frac{p(y|H_0)}{p(y|H_1)} = \frac{p(y|\theta = \theta_0)}{\int_{\Omega_{\theta|H_1}} p(y|H_1, \theta) p(\theta|H_1) d\theta}$ and if it is incredibly small or large then the H_1 or H_0 might be true respectively. An analyst can calculate the $p(H_0|y)$ or $p(H_1|y)$ leading to direct interpretation of the hypothesis given the observed data via

$$\frac{p(H_0|y)}{p(H_1|y)} = BF_{01} \frac{p(H_0)}{p(H_1)}$$

The posterior probability of θ given the observations and the model assumption for the model M_i is

$$p(\theta|y, M_i) = \frac{p(y|\theta, M_i) p(\theta|M_i)}{p(y|M_i)} \propto p(y|\theta, M_i) p(\theta|M_i)$$

Eventually, the general concept of model comparison is implemented via the measurement of the Bayesian factor. Bayesian factor is sometimes very hard to calculate by hand; as a result, various simulation methods are conducted. The problem with the Bayesian factor is that it involves the prior information in the integral, but the prior distribution is less influential when the sample size is big enough. In multivariate problems Bayes factor is impossible to calculate by hand; nevertheless, various simulation methods are conducted to estimate the Bayesian factor. Apart from utilizing the Bayesian factor framework, there is another tool that is based on the predictive function[18,22,23]. In practice, it compares the predictive values from the predictive posterior distribution with the observed ones.

Before, proceeding to computational – simulation methods, it is worth mentioning some asymptotic properties of the random variable θ given the observations. To begin with, as the number of observations is tended to infinite the influence of the prior distribution tends to be minimum as a result the outcome is similar to the frequentist approach. Apart from if the posterior distribution is

unimodal and symmetric the posterior distribution θ given y approximates the normal distribution[18,22,23]. Let assume that the posterior distribution is centered by the posterior mode which is calculated by computation methods. After taking the Taylor series around the posterior mode $\hat{\theta}$, the log posterior distribution is written as a approximation of

$$\log p(\theta|y) \approx \log p(\hat{\theta}\ |y) + [\frac{d \log p(\theta\ |y)}{d\theta}]_{\hat{\theta}} + \frac{1}{2}(\theta - \hat{\theta})[\frac{d^2 \log p(\theta\ |y)}{d\theta^2}]_{\hat{\theta}}(\theta - \hat{\theta})^T$$

The next Taylor tale after the quadratic form tends basically to zero for θ close to $\hat{\theta}$ and n large (where n is the length of the sample). Then the $[\frac{d \log p(\theta\ |y)}{d\theta}]_{\hat{\theta}}$ is zero to the posterior mode. Consequently,

$$\log p(\theta|y) \approx \log p(\hat{\theta}\ |y) + \frac{1}{2}(\theta - \hat{\theta})[\frac{d^2 \log p(\theta\ |y)}{d\theta^2}]_{\hat{\theta}}(\theta - \hat{\theta})^T$$

So, the if the $\log p(\theta|y)$ is considered as function of θ then

$$\log p(\theta|y) \propto \frac{1}{2}(\theta - \hat{\theta})[\frac{d^2 \log p(\theta\ |y)}{d\theta^2}]_{\hat{\theta}}(\theta - \hat{\theta})^T$$

And after the exponent of both terms the posterior odds is analogous to

$$p(\theta|y) \propto \exp(-\frac{1}{2}(\theta - \hat{\theta})I(\hat{\theta})(\theta - \hat{\theta})^T)$$

Where $I(\hat{\theta})$ is the estimated information matrix in $\hat{\theta}$ which is equals to $-[\frac{d^2 \log p(\theta\ |y)}{d\theta^2}]_{\hat{\theta}}$. This relation basically indicates that the posterior distribution is analogous to normal distribution.

Eventually, the random variable of θ given y is approximated by the

$$\theta|\ y\ \sim \mathrm{N}(\hat{\theta}, \hat{I}^{-1}(\theta))$$

This result has numerous applications one of them is the approximation of the posterior distribution when it is difficult to be computed or simulated. Nevertheless, an analyst should check if the posterior mode $\hat{\theta}$ is inside the parameter space because if it's not the information matrix might not be positively defined. Another application is that it provides starting values for simulation methods. Eventually the most important is that for large datasets the influence of the prior distribution is almost zero. Consequently, the Bayesian posterior inference is similar to the frequentist inference when the sample is large enough regardless of the prior distribution [18,22,23].

3.1.2 Bayesian Simulation

More complicated problems are necessary to be solved leading to the soaring complexity of the posterior distribution. In practice as the problems gain more intricacy, more cumbersome algebra is needed for the calculation of posterior distribution and its centrality and variability metrics. Furthermore, in the most complex cases even the posterior distribution is not available and only a non-normalized analogous to posterior distribution is available. In other words,

$p(\theta|y) \propto p(y|\theta)p(\theta) = h(\theta)$, the division of the posterior probability distribution by the non-normalized $h(\theta)$ is equal to an invariable value c which is equal to the $p(y) = \int_\theta p(y|\theta)p(\theta)\, d\theta$ that is pretty much impossible for someone to calculate it. Therefore, in Metropolis Hastings simulation paradigm or the accept reject method only the non-normalized posterior is essential for the simulations[18,24]. Virtually, as it was mentioned the simulations of posterior probability distribution are necessary for the centrality and variability metrics which those metrics are calculated via solving cumbersome integrals. Thus, the simulation framework revolves around solving complicated integrals and the mathematical background is based on numerical approximations and limit theorems [24]. For instance, the $E(g(\theta)|y)$ which the mean of $g(\theta)|y$ or the average of $g(\theta)$ for the random variable $\theta|y$. This theoretical mean is equal to

$$E(g(\theta)|y) = \int_\theta g(\theta)\, p(\theta|y) d\theta$$

This quantity is calculated either theoretically which is impossible in advanced Bayesian framework or numerically either using deterministic methods or stochastic ones like Monte Carlo[24]. Some deterministic numerical integration methods for one-dimension problems are rectangle, trapezoidal, Simpson's rule and more, those rules involve points to estimate the integral[24]. The stochastic integration also called Monte Carlo requires numerous simulated values from the targeted distribution in order to approximate the integral using a form of sample mean. This sample mean concept derives from the Law of Large Numbers which states that let assume $X_1, X_2, \dots, X_n$ random independent and identical variables with mean $E(X_i) = \mu$ and variance $VAR(X_i) = \sigma^2$ for each $i = 1,2, \dots, n$. If $\mu_n = \frac{\sum_{i=1}^n X_i}{n}$ then $\frac{\sqrt{n}}{\sigma}(\mu_n - \mu)$ is approximately distributed to $N(0,1)$ as n tends to infinite. This indicated that μ_n approximates the theoretical μ and the $\frac{\sigma}{\sqrt{n}}$ should be small in order to get more accurate approximations. But how this theorem is connected to the numerical integration of $E(g(\theta)|y) = \int_\theta g(\theta)\, p(\theta|y) d\theta$, let assume that $g(\theta_1|y), g(\theta_2|y), \dots, g(\theta_n|y)$ are independent and identical variables which are independently drawn from the posterior probability distribution. Then the $E\big(g(\theta_i|y)\big) = E(g(\theta)|y) = \int_\theta g(\theta)\, p(\theta|y) d\theta$ for all $i = 1,2, \dots, n$ then using the Law of Large Number theorem the

$$\frac{\sum_{i=1}^n g(\theta_i|y)}{n} \approx E(g(\theta)|y) = \int_\theta g(\theta)\, p(\theta|y) d\theta$$

Therefore, the only issue is to acquire or simulate large number of $\theta_i|y$ from the posterior distribution $p(\theta|y)$ in order to calculate those types of integrals[24]. This concept of approximations is applied to multidimensional problems leading to the definition of a specific problem of how those simulations are generated either to the univariate or multidimension framework given the posterior probability distribution is fully or partially known (without the normalizing constant). Apart from the estimation of centrality and variability methods of the

posterior probability distribution, the simulation of the predicted values is an important issue. Basically, one predicted value $\tilde{y}$ is simulated from $p(\tilde{y}|\theta)$ supposing a simulated value of θ. In other words in order to acquire the predicted posterior distribution first a simulated value θ from $p(\theta|y)$ is required and then this simulated values is used as a condition to simulate a value from $p(\tilde{y}|\theta)$ [18,22,23].

To begin with the most popular and easiest simulation method is Inversion [24]. This method takes the advantage of a property of the probability theory which states that if $X \sim F$ where X is random variable and F it's cumulative function the $F(X) \sim U(0,1)$ where $U(0,1)$ is the uniform distribution. Then if the inverse cumulative function F^{-1} exists then $X \sim F^{-1}(U(0,1))$. Basically, first a value u from uniform distribution is generated and then the $F^{-1}(u)$ is a random value from the distribution of X continuous random variable. For the discreet situation $F^{-1}(u) = \min\{x|F(x) \geq u\}$. In more plain words first generate a u value then the desirable x is that which satisfies the $F(x-1) < u \leq F(x)$. Eventually, this method is very plain in the continuous situations where F^{-1} is easily calculated and in all the discreet situations where just basically the probability distribution is needed. Nevertheless, this method is hard to use when the posterior probability distribution has complex form and when the problem is not univariate. The next method is called Accept – Reject and it basically solves the problem of the posterior complexity [24].This method simulates from another distribution and only a non-normalized form of desired distribution is necessary. So, basically a non-normalized form of posterior distribution is needed and furthermore the generated value is originated from another distribution with at least the same parameter space. Nevertheless, this simulation method suffers from the same multidimensional problem. Initially choose a density function $g(x)$ whose parameter space it's at least as the targeted distribution $f(x)$ parameter space. Then find the $c = \max(\frac{f(x)}{g(x)})$ each value of x belonging to f's parametric space. This c is calculated using the Newton -Raphson method. Then generate a value x from $g(x)$ and an independent value u from uniform distribution. If

$$u \leq \frac{f(x)}{cg(x)}$$

Then keep x and regenerate a x and an u value or discard this x value and just regenerate a x and an u value . The c value is equal to $\frac{1}{p(x\ accepted)}$; as a result, the minimum c is required. Consequently, the difficulty of the method is finding a $g(x)$ which minimizes the maximum of $\frac{f(x)}{g(x)}$. Furthermore, a non-normalized function of $f(x)$ is at least necessary. In the concept of the posterior probability distribution only the $h(\theta)$ which is calculated from $p(\theta|y) \propto p(y|\theta)p(\theta) = h(\theta)$ and a candidate generator $g(\theta)$ is needed to be acquired. Eventually, one final decent univariate simulation method worth mentioning is importance sampling. This importance sampling method generates values for the desirable distribution from another distribution $g(x)$ like the Accept-Reject method [24]. The space which g is defined is at least the same as the targeted distribution $f(x)$. The general concept of importance sampling is let assume that μ is the desirable average.

$$\mu = \int \varphi(x) \frac{f(x)}{g(x)} g(x)dx = E(\varphi(x)w(x))$$

Where $w(x) = \frac{f(x)}{g(x)}$ with $E\big(w(x)\big) = 1,$ is known as the importance weight and g is the importance density. Therefore, they are two estimators for μ

$$\mu_n = \frac{\sum_{i=1}^{n} \varphi(x_i)w(x_i)}{n}$$

And

$$\mu'_n = \frac{\sum_{i=1}^{n} \varphi(x_i)w(x_i)}{\sum_{i=1}^{n} w(x_i)}$$

Where x_i is generated from the importance density $g(x)$.The first estimator μ_n is unbiased and it needs the full formula of $f(x)$ and $g(x)$. The second is biased estimator but it becomes unbiased as n tends to infinity, requiring only a non-normalized formula of $f(x)$ and $g(x)$. The second restriction for $g(x)$ is that the variance of the importance weight is finite. The second estimator has lower variance than the first one when the $g(x)$ is a function which $\varphi(x)\frac{f(x)}{g(x)}$ is approximate constant. So, in practice using a computing software, an analyst can choose various $g(x)$ and plot the $\varphi(x)\frac{f(x)}{g(x)}$ with the aim of observing if this quantity is stable or just, they can experiment with various $g(x)$ in order to check the $\varphi(x)\frac{f(x)}{g(x)}$. Again, this method is very versatile, but it suffers from the multidimensional problem; namely, its purpose is only for univariate problems. Another relative topic about the Monte Carlo method is to use of estimators with low variance to have quick more accurate estimates [24]. One can use consistent estimators that are unbiased and with low variance as n tends to infinity. Or they can use unbiased estimators with the lowest variance. Nevertheless, those estimators are impossible to use in the Bayesian concept because they are hard to derive. Last but not least, some distributions are derived from a complication of another distribution. For example, the Binomial distribution is the sum of Bernoulli ones. Sometimes it is easier to simulate several Bernoulli ones and then take a sum rather than the direct simulation of a Binomial. Also, one can take into consideration other variable transformations to calculate the desired estimator. Those methods are very useful for univariate problems but virtually most problems are multidimensional. Those issues are solved by simulation methods such as Gibbs sampler and Metropolis Hasting with the first one being a special case of the latter.

Gibbs sampling method is a special case of the Metropolis Hasting algorithm and its purpose is to simulate parameter values from the posterior probability distribution[18,24]. This is achieved by not simulating from the joint posterior probability distribution but from the conditional posterior distributions. This specific concept makes Gibbs sampling a powerful method to acquire simulations from the joint posterior distribution. Particularly let's assume that blocks of the parameter $\theta = (\theta_1, \theta_2, .., \theta_d)$ are necessary to be simulated where the dimension of θ_j is at least 1 for $j = 1,2, \dots, d$, then the first step is to acquire the

conditional posterior distributions. The conditional posterior distribution for the θ_j block is $p(\theta_j|\theta_{-j}, y) \propto p(\theta|y) \propto p(y|\theta)p(\theta) = h_j(\theta_j)$ for $j = 1, 2, , \dots, d$ where $\theta_{-j} = (\theta_1, \theta_2, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_d)$. Therefore, the normalized version of h_j or just $p(\theta_j|\theta_{-j}, y)$ is necessary for each $j = 1, 2, , \dots, d$. The calculation of $p(\theta_j|\theta_{-j}, y)$ is conducted via the conditional conjugacy. The conditional conjugacy is the same property as the conjugate property which states that the $p(\theta_j|\theta_{-j}, y)$ belongs to the same family as the $p(\theta_j)$. In other words, one can write the $p(y|\theta)p(\theta)$ and then take the terms which are function of θ_j assuming θ_{-j} as fixed , then using the conjugacy property or by normalizing the $h_j(\theta_j)$ calculates the $p(\theta_j|\theta_{-j}, y)$. Then an analyst assigns some initial values $\theta^0 = ({\theta_1}^0, {\theta_2}^0, .., {\theta_d}^0)$ and first ${\theta_1}^1$ is simulated from the $p(\theta_1|\theta_{-1}^0, y)$, then ${\theta_2}^1$ from the $p(\theta_2|{\theta_1}^1, {\theta_3}^0, .., {\theta_d}^0, y)$. .. and for the last ${\theta_d}^1$ is simulated from the $p(\theta_d|\theta_{-d}^1, y)$. Then the first round of simulation has been conducted. The second round is similarly conducted. Let assume that it is time for the ith round, the ${\theta_j}^i$ is simulated from the

$$p(\theta_j|{\theta_1}^i, {\theta_2}^i, \dots, \theta_{j-1}^i, \theta_{j+1}^{i-1}.., {\theta_d}^{i-1}, y)$$

So, the ${\theta_j}^i$ is simulated given the most updated values of θ. A large number of simulations are conducted in order to achieve convergence and after the so-called burn in period the next simulations are drawn for inference and summarization.

As it is stressed Gibbs sampling is a specific case of the Metropolis – Hasting Algorithms [18,24]. The Gibbs sampling method works well when the conditional posterior distributions are known; namely, one can simulate the desirable values from those conditional posterior distributions. The Metropolis-Hastings has one particularity it generates simulations from another distribution; as a result, it handles problems when the conditional posterior distribution is too complicated to simulate from. Especially it only utilizes the non-normalized aspect of the conditional posterior distribution. Nevertheless, the general concept which is the endeavor to simulate values from the joint posterior distribution through the conditional posterior distribution remains the same. The general algorithm is, let say that the ${\theta_1}^i, {\theta_2}^i, .., {\theta_d}^i$ have been simulated where ${\theta_1}^i$ is a group of or just one simulated value of the first set of parameters in the i iteration. Then the ${\theta_1}^{i+1}$ is the next simulated value for the upcoming round $i + 1$. One possible ${\theta_1}^{i+1}$ which is written as θ_1^{pos} is generated by a specific or sometimes arbitrary distribution $q(\theta_1|{\theta_1}^i, {\theta_2}^i, .., {\theta_d}^i)$ then the ${\theta_1}^{i+1}$ is equal to θ_1^{pos} with probability p or it is equal to ${\theta_1}^i$ with probability $1 - p$. This probability p is defined as

$$p(\theta_1^{pos}, {\theta_1}^i) = \min\{1, \frac{p(\theta_1^{pos}|{\theta_2}^i, .., {\theta_d}^i)}{p(\theta_1^i|{\theta_2}^i, .., {\theta_d}^i)} \frac{q(\theta_1^i|{\theta_1}^{pos}, {\theta_2}^i, .., {\theta_d}^i)}{q(\theta_1^{pos}|{\theta_1}^i, {\theta_2}^i, .., {\theta_d}^i)}\}$$

For θ_j^{i+1} the same algorithm is used first a value θ_j^{pos} is generated from the $q(\theta_j|{\theta_1}^{i+1}, {\theta_2}^{i+1}, \dots \theta_{j-1}^{i+1}, {\theta_j}^i, \theta_{j+1}^i .., {\theta_d}^i)$ and then the $p(\theta_j^{pos}, {\theta_j}^i)$ is equal to

$$\min\{1, \frac{p(\theta_j^{pos}|{\theta_1}^{i+1}, {\theta_2}^{i+1}, \dots \theta_{j-1}^{i+1}, \theta_{j+1}^i .., {\theta_d}^i)}{p(\theta_j^i|{\theta_1}^{i+1}, {\theta_2}^{i+1}, \dots \theta_{j-1}^{i+1}, \theta_{j+1}^i .., {\theta_d}^i)} \frac{q({\theta_j}^i|{\theta_1}^{i+1}, {\theta_2}^{i+1}, \dots \theta_{j-1}^{i+1}, \theta_j^{pos}, \theta_{j+1}^i .., {\theta_d}^i)}{q(\theta_j^{pos}|{\theta_1}^{i+1}, {\theta_2}^{i+1}, \dots \theta_{j-1}^{i+1}, {\theta_j}^i, \theta_{j+1}^i .., {\theta_d}^i)}\}$$

Some notable comments are that first the generator q is possible to be an arbitrary choice which implies that every possible generator leads to convergence. If the generator q is equal to the conditional posterior distribution, then the probability of acceptance is one leading to the implementation of Gibbs sampling. Furthermore , the completed knowledge of the posterior is not necessary, only a non-normalized posterior analogous function is needed. One prominent generator because its simplicity is a normal distribution with mean equal to the previous θ_κ^i value and known variance. This is called Random walk Metropolis algorithm [18,24] and it is appropriate and simple because of the symmetry which normal distribution provides. Specifically, for the θ_j^{i+1} simulation the

$$p\left(\theta_j^{pos}, {\theta_j}^i\right) = \min\{1, \frac{p(\theta_j^{pos}|{\theta_1}^{i+1}, {\theta_2}^{i+1}, \dots \theta_{j-1}^{i+1}, \theta_{j+1}^{i} .., {\theta_d}^i)}{p(\theta_j^i|{\theta_1}^{i+1}, {\theta_2}^{i+1}, \dots \theta_{j-1}^{i+1}, \theta_{j+1}^{i} .., {\theta_d}^i)}\}$$

where $q({\theta_j}|{\theta_1}^{i+1}, {\theta_2}^{i+1}, \dots \theta_{j-1}^{i+1}, {\theta_j}^i, \theta_{j+1}^i .., {\theta_d}^i)$ is $N\left({\theta_j}^i, \Sigma\right)$ with known Σ. The covariance matrix is tuned by the acceptance rate[24]. When the acceptance rate is low the algorithm convergences fast and when the acceptance rate is bigger the convergence is lower. In practice when the dimension of the target parameter is one or two 0.5 acceptance rate is suggested and when the parameter dimension is big at least 0.25 recommended[24]. Therefore, when the Random walk Metropolis algorithm is used the fixed parameters are tuned by the acceptance probability. Eventually, the choice q may be different for every θ_κ $\kappa = 1,2, \dots, d$ which leads to the inference that a mixed type of Metropolis Algorithms is frequently utilized.

The upcoming issue after the implementation of the previous algorithms is the assessment of convergence[24]. Practically the problem is to assess if the generated values have been converged to the desired target distribution. This is done by collecting the simulated values from different starting points. For each starting point a chain of simulations in the form of sequence is conducted. Practically, the strategy is to acquire several chains to check the convergence. Particularly every chain should converge to the same distribution, therefore, every chain should be stationary and be near other chains. For example, two chains are stationary but when they are plotted together, they seem to converge in different distributions; as a result, the mixing of the two chains is not successful. One another problem is that in the Metropolis Hasting algorithm, it is obvious that on average every value is dependent on the previous one. So, there is autocorrelation, this indicates that the autocorrelation of different lags should be calculated. To tackle those issues, the researchers have pointed out that the first half of each chain should be deleted because the first half values are somehow correlated to the starting points and because convergence might not be achieved[24]. So, after the burn-in period the autocorrelation of each halved chain should be calculated and choose a specific lag or k which $cov\left(\theta^i, \theta^{i+k}\right)$ is close to zero or just decreasing. So, after the choice of every kth values of each simulation, an analyst can break them in further groups (or just if they have enough chains, they should keep it as it is) with aim of checking both the mixing and stationary convergence. This is conducted via the calculation of the scale reduction factor or just $\hat{R}$ which

is function of between and within variance of the chains or remaining groups of simulations [18,24]. Let assume that there are m chains and n simulations in each chain. Also ψ_{ij} is the i value of the j chain where $i = 1,2, \dots n$ and $j = 1,2, \dots, m$. The Between variance is described as

$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{.j} - \bar{\psi}_{..})^2$where $\bar{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}$ and $\bar{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{.j}$. The Within variance is derived from the $W = \frac{1}{m}\sum_{j=1}^{m}s_j^2$ where $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{ij} - \bar{\psi}_{.j})^2$. Theoretically if the convergence is achieved the Between variance divided by n should be close to zero. With those quantities the variance of the $\psi|y$ is calculated through the formula $\frac{n-1}{n}W + \frac{1}{n}B$. The scale reduction factor is

$$\hat{R} = \sqrt{\frac{\widehat{Var(\psi|y)}}{W}} = \sqrt{\frac{n-1}{n} + \frac{B}{nW}}$$

This $\hat{R}$ is equal to 1 for n going to infinite. So, a appropriate convergence indicated a $\hat{R}$ between 1 and 1.1 [18,24]. To sum up, the parameters of prior distribution or the generator function are determined by the acceptance probability and the number of n the number of each chain (the number of chains are at least two) are determined by the correlation and $\hat{R}$.

In the most cases the problems which are arises are related with various variables, one of them is the desirable outcome (Y) and the others are explanatory variables which are trying to explain the variability or the fluctuation of the outcome. In those situations, an analyst can implement a generalized linear models or other models like survival and more [25]. Those models have in common a construction of likelihood function and a relation which links the outcome variable and the explanatory variables. In the general situation let assume that there are n observations, the outcome variable Y and the predictors X. Also, there are some parameters which they are trying to measure the association between the outcome and the explanatory. So, the purpose of statistics either frequentist or Bayesian is to estimate those parameters. Let assume that the parameters are $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ and there is a model which describes the relation between Y and X. This general model is described as $p(y_i|\beta, x_i)$ for each $i = 1,2, \dots n$, then in the frequentist approach those β frequently are calculated by the maximation of the likelihood function $\prod_{i=1}^{n}p(y_i|\beta, x_i)$. In the Bayesian approach those parameters are variables and they have a prior distribution lets say $p(\beta)$. The main purpose of the Bayesian approach is to estimate the posterior distribution of the desirable parameters $p(\beta|y, x)$. This posterior probability distribution is described as

$$p(\beta|y, x) \propto \prod_{i=1}^{n} p(y_i|\beta, x_i)p(\beta)$$

Practically, only the non-normalized version of the posterior distribution is known; therefore, the employment of Metropolis-Hasting algorithm is reasonable.

Let consider a theoretical problem which a logistic regression is suggested. Furthermore, let presume $y_i \sim B(n_i, p_i)$ which is the target variable and $x_i^T = (1, x_{1i}, x_{2i}, \dots, x_{pi})$ are the explanatory variables for the ith observation. The binary logistic regression assumes that

$$logit\ p_i = \log \frac{p_i}{1 - p_i} = \eta_i = x_i^T \beta$$

Where $\beta = (\beta_0, \beta_1, \dots, \beta_p)$. The likelihood the model is

$$p(y|\beta, x) = \prod_{i=1}^{n} \frac{\exp(y_i x_i^T \beta)}{1 + \exp(x_i^T \beta)^{\eta_i}}$$

The prior distribution of the β is multivariate normal distribution $N_p(\mu_0, \Sigma_0)$ [25] where μ_0 and Σ_0 are chosen by some previous information or they can be arbitrary or they are chosen in order to satisfy other conditions like the probability of acceptance. So the posterior distribution is written as

$$p(\beta|y, x) \propto \prod_{i=1}^{n} \frac{\exp(y_i x_i^T \beta)}{1 + \exp(x_i^T \beta)^{n_i}} \exp\left(-\frac{(\beta - \mu_0)^T {\Sigma_0}^{-1} (\beta - \mu_0)}{2}\right)$$

Because the precise form of the posterior probability distribution is hard to be calculated, one can proceed to Metropolis Hasting. The β are generated either jointly or one by one. There are several algorithms which can conduct this concept but one of the simplest is the random walk. The β are simulated by the Metropolis Hasting theorem and then they are summarized in order to interpret the nature of the explanatory variables. The significant of each variable is checked either by calculating the Bayesian factors [26,27] which are estimated through Monte Carlo. Particularly, first simulations from the prior distributions are conducted and then for each simulation the likelihood of the given hypothesis is calculated. Finally the sample mean of every estimated likelihood is taken. Other measures which are related to the frequentists p-values are probability of direction(PH) [26,27] which is the proportion of the simulations having the same sign as the simulated sample median , region of practical equivalence (ROPE) [26,27] which it checks the proportion of the central interval that is included in the ROPE interval and central interval. Every other problem follows a similar concept, first the likelihood is calculated then some prior assumptions are done and then an analyst must choose which simulation algorithm is better.

3.2 Weibull Survival Model

The typical Cox regression model does not assume a specific form about the baseline hazard function or just the hazard function [4]. Specifically, Cox models relates the hazard function to baseline hazard function whose form is unknown. The parametric proportional models , additionally to the proportional property, assumes that the hazard or in practice the baseline hazard function has a specific formula. This extra property leads to more precise coefficient parameter estimations or just the cox standard errors are bigger than the parametric ones.

The parametric proportional hazard model which is analyzed is the Weibull model [4] which has the prominent property of monotonous or sometimes steady hazard function. The survival time has a density function $f(t)$, survival function $S(t) = 1 - \int_0^t \int f(u)du$, a hazard function $h(t) = \frac{f(t)}{S(t)}$ and the cumulative hazard function $H(t) = -\log S(t)$. The hazard function under the Weibull assumption has a specific form

$$h(t) = \lambda\gamma t^{\gamma-1}$$

Where $t \geq 0$ and λ, γ are both grater than zero. If $\gamma = 1$ (exponential distribution), the hazard function is steady and equal to λ and if the $\gamma \neq 1$ the hazard function is monotonous. The shape and scale parameter is γ and the λ respectively.Therefore, $f(t) = \lambda\gamma t^{\gamma-1} \exp(-\lambda t^{\gamma})$ and $S(t) = \exp(-\lambda t^{\gamma})$. Eventually the pth percentile of survival time $t(p) = \{\frac{1}{\lambda}\log\left(\frac{100}{100-p}\right)\}^{1/\gamma}$. The scale and shape parameters λ, γ are estimated via the likelihood function. In the frequentist approach the likelihood is maximized and in the Bayesian approach those two parameters are estimated through simulations via assigning two prior distributions. Specifically, the likelihood function for independent right censoring is

$$\prod_{i=1}^{n} h_i(t)^{\delta_i} S_i(t) = \prod_{i=1}^{n} (\lambda\gamma t_i^{\gamma-1})^{\delta_i} \exp(-\lambda t_i^{\gamma})$$

Where n are the number of observations and δ_i is equal to 1 if an observation is censored and 0 if not.

Let assume that there are measure other variables $X_1, X_2, \dots, X_p$ than time and censoring. Therefore, for every individual ith for $i = 1,2,.,n$ t_i, $\delta_i, x_{i1}, x_{i2}, \dots, x_{ip}$ are observed. Under the proportional models without giving a specific form to the base hazard function; namely, the Cox model the hazard function for a specific individual is

$$h_i(t) = \exp\left(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) h_0(t)$$

Where $\frac{h_i(t)}{h_0(t)}$ is independent of time and the $h_0(t)$ does now has a specific form. Under the Weibull assumption the $h_0(t)$ adapt a defined formula which is $\lambda\gamma t^{\gamma-1}$. Consequently, the hazard function of the parametrically proportional hazard is

$$h_i(t) = \exp\left(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \lambda\gamma t^{\gamma-1}$$

Also, the survival, cumulative hazard and density function has a particular form. The survival function is defined as

$$S_i(t) = \exp\left(-\exp(\beta' x_i)\, \lambda t^{\gamma}\right)$$

Where $\beta' x_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ [4]. So, the likelihood function is equal to $\prod_{i=1}^{n} h_i(t)^{\delta_i} S_i(t)$ and if the $h_i(t)$ and $S_i(t)$ are replaced. Then the likelihood is

$$\prod_{i=1}^{n}[\exp(\beta' x_i)\lambda\gamma t^{\gamma-1}]^{\delta_i}\exp\left(-\exp(\beta' x_i)\,\lambda t^{\gamma}\right)$$

Finally, the quantile $t(p)$ is equal to $\{\frac{1}{\lambda\exp(\beta' x_i)}\log\left(\frac{100}{100-p}\right)\}^{1/\gamma}$. The parameters are estimated either using the frequentist approach (maximization) or utilizing the Bayesian concept of simulating those parameters and then summarizing them. So, after having those estimated parameters $\hat{\beta}$, $\hat{\lambda}$ and $\hat{\gamma}$ either by maximizing the likelihood or just take the mean of the simulations. The estimated model is $h(t) = \exp(\hat{\beta}' x)\hat{\lambda}\hat{\gamma}t^{\hat{\gamma}-1}$ and it is imperative to assess the model validity. The most prominent model checking residuals are Cox Schnell[4], martingale residuals or deviance residuals. Cox-Snell residuals are

$$r_{C_i} = \hat{H}_i(t_i) = -\log\hat{S}_i\,(t_i) = \exp(\hat{\beta}' x_i)\hat{\lambda}t_i^{\hat{\gamma}}$$

The Cox-Snell residuals given the theory; they follow unit exponential distribution. The martingale residuals[4] are derived from the Cox-Snell residuals their form is

$$r_{M_i} = \delta_i - r_{C_i}$$

Unusual large values of martingale residuals indicate that the specific observations are not well fitted. Also, they are not symmetrically distributed. Next there are the deviance residuals[4] which they derived from

$$r_{M_i} = sign(r_{M_i})\sqrt{-2(r_{M_i} + \delta_i\log(\delta_i - r_{M_i}))}$$

Deviance residuals are an endeavor to make the martingale residuals symmetric around zero. In the frequentists approach the Schoenfeld residuals are calculated using the maximum likelihood estimators or in general the maximum likelihood approach. This is not applied to Bayesian survival because the estimators are derived from the sample mean of the simulation and they tend to be the same with the maximum likelihood estimators if the sample is large enough. In practice, I think that the best strategy to check the proportionality is to introduce dependent coefficients and variable selection methods or just one can use the corresponding frequentists p-value methods to Bayesian framework like Bayesian factors to check if the coefficient of each variable multiplied by the survival time is 'Bayesian significant '.

3.3 Bayesian Weibull survival model.

As it was stressed, to find a non-normalized fraction of the posterior distribution of the parameters or the same posterior distribution itself. The likelihood function of the model and the prior distribution of the parameters are eventually necessary. So, let assume that there are measures of other variables $X_1, X_2, \dots, X_p$ than time and censoring. Therefore, for every individual ith for $i = 1,2,.,n$ t_i, $\delta_i, x_{i1}, x_{i2}, \dots, x_{ip}$ are observed where t_i and δ_i is the survival time and the censored indicator indicating 1 when the observation is censored and 0 elsewhere. If all the variables are included in the model, then $p + 2$ parameters are

essential to construct the Bayesian Weibull model [28]. First the scale and the shape parameters λ, γ and the $\beta' = (\beta_1, \beta_2, \dots, \beta_p)$. The likelihood of the observations given the parameters is

$$p(x, t, \delta|\lambda, \gamma, \beta') = \prod_{i=1}^{n} [\exp(\beta' x_i)\lambda\gamma t^{\gamma-1}]^{\delta_i} \exp(-\exp(\beta' x_i)\lambda t^{\gamma})$$

The posterior distribution of λ, γ are Gamma distribution with parameters a_λ,c_λ and a_γ,c_γ shape and scale parameters respectively where mean of λ is equal to $a_\lambda * c_\lambda$. The vector β' is following multivariate normal distribution $N_p(\mu_0, \Sigma_0)$ with known μ_0, Σ_0. Those a_λ,c_λ, a_γ,c_γ, μ_0, Σ_0 are either arbitrary or are tuned to satisfy some conditions like the acceptance probability or they are derived from previous information. For example, μ_0 is 0 because it indicated the initial hypothesis. a_λ,c_λ, a_γ,c_γ are tuned to have simulations acceptance rate equal to 0.5 or they just create high enough variance. Previously information is derived from previous studies and more. Eventually, sensitivity analysis is frequently conducted to check the sensitivity of the prior information. Therefore, the posterior distribution is analogous to

$$p(\lambda, \gamma, \beta'|x, t, \delta) \propto$$

$$\prod_{i=1}^{n} [\exp(\beta' x_i)\lambda\gamma t^{\gamma-1}]^{\delta_i} \exp(-\exp(\beta' x_i)\lambda t^{\gamma}) \exp\left(-\frac{(\beta-\mu_0)^T {\Sigma_0}^{-1}(\beta-\mu_0)}{2}\right) g(\lambda, a_\lambda, c_\lambda) g(\gamma, a_\gamma, c_\gamma)$$

The conditional posterior distribution of the parameters is

$$p(\beta'|x, t, \delta, \lambda, \gamma) \propto \prod_{i=1}^{n} [\exp(\beta' x_i)]^{\delta_i} \exp(-\exp(\beta' x_i)\lambda t^{\gamma}) \exp\left(-\frac{(\beta-\mu_0)^T {\Sigma_0}^{-1}(\beta-\mu_0)}{2}\right)$$

$$p(\lambda|x, t, \delta, \gamma, \beta') \propto \prod_{i=1}^{n} [\lambda]^{\delta_i} \exp(-\exp(\beta' x_i)\lambda t^{\gamma}) g(\lambda, a_\lambda, c_\lambda)$$

$$p(\gamma|x, t, \delta, \lambda, \beta') \propto \prod_{i=1}^{n} [\gamma t^{\gamma-1}]^{\delta_i} \exp(-\exp(\beta' x_i)\lambda t^{\gamma}) g(\gamma, a_\gamma, c_\gamma)$$

Those conditional posterior distribution can further be simplified. Then the Metropolis-Hasting algorithm with random walk is applied. The parameters of the generators are simply tunned by the accept probability criteria. If the normal generator for λ and γ produce negative values then the probability of acceptance is zero. After finishing the simulation phase with respect of the other convergent criteria, the sample means of the simulations are the estimators of the parameters. Therefore, the model checking is conducted via calculation of the Cox-Snell, martingale and deviance residual. For instance the Cox-Snell residuals are derived from

$$r_{C_i} = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i) = \exp(\hat{\beta}' x_i)\hat{\lambda} {t_i}^{\hat{\gamma}}$$

In conclusion, only the necessary basics of Bayesian statistics are written and there are far more topics that have not been represented like the hierarchical concept, variable selection, predictive model checking, decision theory, and more.

Furthermore, there are more parametric models than the Weibull model, some of them are called accelerated time failure models, and the Weibull is also part of them. Those accelerated models describe the relation between the time and the covariates like the generalized linear models. In practice, there is a linear relation between the logarithm of the survival times and the covariates and not the hazard function.

4 Methodology

4.1 The models

4.1.1 Introduction

This master thesis is about first deploying a Bayesian parametric competing risk model with missing event types under MAR assumption with independent right censoring. Second, the results of the Bayesian method are compared with one frequentist approach. Finally, both approaches are compared with the simulation data model. In general, the Bayesian parametric competing risk model with missing cause of failure is a mixture of two methodologies. The first methodology is about employing a Bayesian parametric Weibull competing risk model in a complete dataset using the Metropolis Hasting Algorithm [29]. The second methodology is about Bayesian imputation methods for missing data [17] which generally implies how to develop Bayesian models and use Markov Chain Monte Carlo methods to impute missing values. So, under one Metropolis Hasting iteration firstly the missing values are imputed, and then using a full dataset the Bayesian Weibull competing risk model is updated. The previous procedure is continued for numerous steps until the convergence is achieved. The frequentist estimation method is exactly the method which is described by Bakoyiannis and others [16], this model is the typical Cox competing risk model with the difference that the observations with missing event types have weights. It is a weighted Cox regression that is deployed for each event type. The observations with missing event types are duplicated according to the number of possible events. For instance, under MAR assumption and with two event types the missing observations are doubled, in the first bunch the first event indicator is imputed, and their weights are related to the probability of the first event being the real one and in the second bunch the second event indicator is assigned with weights which are equal to the probability of the second event being caused. Those probabilities are modeled in the same way as Lu and Tsiatis [15] and Bakoyannis and others [3,9] proposed. Another possible frequentist method is the multiple imputation method estimation which was described by Lu and Tsiatis [15] and Bakoyannis and others [9]. This multiple imputation method practically generates numerous datasets and then the missing indicator is imputed with predicted values from a probability model. Then after the creation of several complete datasets, the normal competing risk analysis is implemented and then the results are aggregated in the same way as every typical multiple imputation analysis. One issue that arises in every proposed method is the modeling of the event probability model which plays a crucial role because if this model is biased then each of those 3 approaches is highly negatively influenced by this misstep. To address this problem, several variable model strategies are applied such as fractional polynomial, splines, interactions, quadratic, cubic, and higher powers modeling, and far more [9,15,16].In the Bayesian approach firstly the missing observations are imputed with predictions from posterior predicted distribution (Imputation step) each turn [17] and then the probability model is updated according to the Bayesian framework. Apart from the requested Bayesian approach, the weighting approach is used for comparison because this approach has shown empirically more efficient results than the multiple imputation method [16].

4.1.2 Bayesian parametric competing risk with missing event type.

A convenient way to construct this model is to break it into pieces. So, one can say that the model is a fusion of two methodologies, the first one is the Bayesian parametric competing risk [29] and the other is the Bayesian handling of missing cause of failure under MAR assumption[17].Let assume that there are n observations, 2 cause of failures, the d_{ki} indicator which is basically equal to 1 if the ith individual has experience the kth event and 0 if it is not $i = 1, 2,3, \dots, n$ and $k = 1, 2$, the time t_i until the experienced event or the end of the survey or a lost of follow up and finally a set of predictors $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ for each individual where p is the number of variables. The censoring is independent right and for now the cause of failure is fully observed. The hazard function that is studied is the cause – specific hazard function. So according to the first chapter, the likelihood function of the competing risk is

$$L = \prod_{i=1}^{n} \prod_{k=1}^{2} [h_k(t_i, x_i)]^{d_{ki}} \, \mathrm{S}_k(t_i, x_i)$$

As it was pointed out for each event type an almost independently survival model is deployed[1,2,3]. In other words, in this instance, there are two models one for each event type. In the first event type all other event types are assumed as censored and the same concept is carried out in the second event type. Therefore, the two likelihoods are

$$L_1 = \prod_{i=1}^{n} [h_1(t_i, x_i)]^{d_{1i}} \, \mathrm{S}_1(t_i, x_i)$$

$$L_2 = \prod_{i=1}^{n} [h_2(t_i, x_i)]^{d_{2i}} \, \mathrm{S}_2(t_i, x_i)$$

The analysis is conducted separately but with the same covariates, the two hazard and survival function are different. Therefore, now let assume that time follows Weibull distribution, the hazard function and the survival function has a specific form. In the general situation without competing risk the hazard function is equal to

$$h(t_i, x_i) = \exp(\beta' x_i) \lambda \gamma {t_i}^{\gamma - 1}$$

and the survival function is equal to

$$S(t_i, x_i) = \exp\left(-\exp(\beta' x_i)\, \lambda {t_i}^{\gamma}\right)$$

Where $\beta' = (\beta_1, \beta_2, \dots, \beta_p)$ are the coefficients of $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and the $\beta' x_i$is equal to $\sum_{j=1}^{p} \beta_1 x_{i1} + \beta_2 x_{i1} + \cdots + \beta_p \, x_{i1}$. Also, λ and γ are the shape and scale parameters respectively. Now, in the presence of competing risks there are two hazards and two corresponding survival functions. The λ_1, γ_1 and λ_2, γ_2 are the scale and the shape parameters of the first and the second event type.

Therefore, the Weibull competing risk is basically two models with one likelihood each. The coefficients of those two models are found by the usual ways such us likelihood maximization and more. The two likelihoods are written as

$$L_1 = \prod_{i=1}^{n} [\exp(\beta_1' x_i)\lambda_1 \gamma_1 t_i{}^{\gamma_1 - 1}]^{d_{1i}} \exp(-\exp(\beta_1' x_i)\lambda_1 t_i{}^{\gamma_1})$$

$$L_2 = \prod_{i=1}^{n} [\exp(\beta_2' x_i)\lambda_2 \gamma_2 t_i{}^{\gamma_2 - 1}]^{d_{2i}} \exp(-\exp(\beta_2' x_i)\lambda_2 t_i{}^{\gamma_2})$$

In conclusion, the parameters of both models are found independently, for instance the parameters of the first and second model are found by the maximization of the first and second likelihood respectively. In other words, the competing risk situations when using cause-specific hazard function is broken down to a specific number of event type of standard non-competing risk analysis.

The Bayesian Weibull competing risk model is virtually a different way to calculate the coefficient estimations. The inference of Bayesian analysis is conducted via the calculation of the posterior distribution. This posterior distribution of the parameters in the multivariate concept almost always is estimated by Monte Carlo simulations. The posterior distribution of the parameters is analogous to the likelihood of the model multiplied by the prior distribution of them. In this particular scenario, the two likelihoods of the two independently models are known, so the only "problem" is to assign prior distribution to the parameters. From the final paragraph of the previous chapter the Bayesian Weibull survival model , the β coefficients are following multivariate normal distribution with known parameters and the shape and scale parameters are following gamma distribution with known parameters. So, the posterior distribution of parameters is analogous to the survival likelihood multiplied by the prior distribution which is written as

$$p(\lambda, \gamma, \beta' | x, t, \delta) \propto$$

$$\prod_{i=1}^{n} [\exp(\beta' x_i)\lambda\gamma t^{\gamma-1}]^{\delta_i} \exp(-\exp(\beta' x_i)\lambda t^{\gamma}) \exp\left(-\frac{(\beta-\mu_0)^T {\Sigma_0}^{-1}(\beta-\mu_0)}{2}\right) g(\lambda, a_\lambda, c_\lambda) g(\gamma, a_\gamma, c_\gamma)$$

Where β' are the coefficients, λ and γ are the scale and shape parameters, a_λ and c_λ are the shape and scale parameters of the λ distribution, a_γ and c_γ are the shape and scale parameters of the γ distribution and finally μ_0, Σ_0 are the mean and matrix variance of the β. There are two posterior distributions in the presence of the competing risk with different parameters each because there virtually two different models – likelihoods. For the first event, the posterior distribution is described as

$$p(\lambda_1, \gamma_1, \beta_1 | x, t, \delta_1) \propto$$

$$\prod_{i=1}^{n} [\exp(\beta_1' x_i)\lambda_1 \gamma_1 t_i{}^{\gamma_1 - 1}]^{d_{1i}} \exp(-\exp(\beta_1' x_i)\lambda_1 t_i{}^{\gamma_1}) \exp\left(-\frac{(\beta_1-\mu_{01})^T {\Sigma_{01}}^{-1}(\beta_1-\mu_{01})}{2}\right) g(\lambda_1, a_{\lambda_1}, c_{\lambda_1}) g(\gamma_1, a_{\gamma_1}, c_{\gamma_1})$$

The same concept is applied for the second event

$$p(\lambda_2, \gamma_2, \beta_2 | x, t, \delta_2) \propto$$

$$\prod_{i=1}^{n} [\exp(\beta_2' x_i) \lambda_2 \gamma_2 t_i^{\gamma_2 - 1}]^{d_{2i}} \exp(-\exp(\beta_2' x_i) \lambda_2 t_i^{\gamma_2}) \exp\left(-\frac{(\beta_2 - \mu_{02})^T \Sigma_{02}^{-1} (\beta_2 - \mu_{02})}{2}\right) g(\lambda_2, a_{\lambda_2}, c_{\lambda_2}) g(\gamma_2, a_{\gamma_2}, c_{\gamma_2})$$

Where the prior distribution of β_1, β_2 is multivariate normal distribution with known mean μ_{01}, μ_{02} and Σ_{01}, Σ_{02} covariance matrix respectively. The prior distribution of λ_1, γ_1, λ_2, γ_2 is gamma with shape and scale parameters a_{λ_1}, a_{γ_1} $a_{\lambda_2}, a_{\gamma_2}$ and $c_{\lambda_1}, c_{\gamma_1}, c_{\lambda_2}, c_{\gamma_2}$ respectively. The way to estimate the parameters is via simulations. The simulation algorithm which is used in this thesis is Metropolis Hasting[29]. The general algorithm of Metropolis Hasting, which has been described in a previous chapter, starts with the assignment of initial values to the parameters then possible candidates are simulated for a specific parameter group from a specific generator. Then those simulations are accepted with a probability of acceptance, or they are rejected. Nevertheless, the dataset must be complete to start the candidate generators in the simulation.

There is a Bayesian method for imputing missing values such as missing event types. This method finds the posterior predictive probability of the missing cause of failure and using the predictive probability imputes the dataset each turn [17]. The predictive posterior probability of the missing event types in the presence of two competing risks is a Bernoulli distribution. The probability of the Bernoulli distribution is derived from the general linear model. Therefore, the concept is to first calculate the probability that is needed to simulate the predicted values. The probability is calculated by the general linear model assuming the existence of its parameters. Then after the imputation of the dataset simulate and update the parameters of the general linear model. The powerful aspect of this method is that in the same iterations of the Metropolis Hasting, both the data imputation and the simulation of Weibull's competing risk parameters are combined. So, the general concept is that in each Metropolis Hasting iteration first the dataset is imputed and then the Weibull competing risk model parameters are simulated. This happens each turn until the end of the iterations. The algorithm for the imputation of missing event types is described as. So, let's assume that there are n observations, n_1 are observed and $n - n_1$ are missing. The first step is to impute the missing events from the posterior predictive probability $p(y_{miss}|y_{obs}, \theta, z)$ and then using the complete non-censored dataset update the θ parameter from the $p(\theta|y_{miss}, y_{obs}, z)$ where y is equal to 1 if the first event has been caused or 0 for the second event (censoring observations are not included) and z all the possible variables including the survival time and their transformations. The posterior predictive probability is basically a Bernoulli distribution with probability p equal to $\frac{\exp(\theta' z)}{1+\exp(\theta' z)}$. Or basically it is derived from the

$$logit\ p = \log\frac{p}{1-p} = \eta = \theta' z$$

Where z is all the possible variables which can predict the event type including the survival time, θ are the coefficients. This general linear model must be constructed with meticulousness and great attention. All possible interactions or fractional polynomials or splines and more should be included. So, after the imputation of the missing event type the parameter θ is simulated using the complete non-

censored dataset. The posterior distribution of the parameter θ is analogous to the likelihood of the general model multiplied by the prior distribution. The prior distribution follows multivariate normal distribution with fixed mean and matrix covariance. The form of the posterior distribution of θ is

$$p(\theta|y,z) \propto \prod \frac{\exp(y_i \theta' z_i)}{1+\exp(\theta' z_i)} \exp\left(-\frac{(\theta-\mu_\theta)^T {\Sigma_\theta}^{-1}(\theta-\mu_\theta)}{2}\right)$$

The μ_θ and Σ_θ is the mean and covariance matrix of the θ's prior distribution. In addition, someone can exploit the asymptotic property of the posterior distribution which has been described in the previous chapter. Particularly, the posterior distribution asymptotically follows the normal multivariate distribution. In other words,

$$\theta|y,z \sim N(\hat{\theta}, \hat{I}^{-1}(\theta))$$

Where $\hat{\theta}$ is the mode of the posterior distribution which is basically the likelihood estimation of the logistic model. The $\hat{I}^{-1}(\theta)$is the estimated inversed information matrix which is equal to the estimated covariance matrix of the parameters. The asymptotic method is used for convenient reasons and because in practice, there is no need to have a cumbersome simulation method for the logistic model. Also, this asymptotic property is valid when the sample size is big enough and has a flat prior distribution. Additionally, with the asymptotic model, there is no need to check the convergence criteria of the logistic model in the end because the coefficients are straightforwardly simulated from the target distribution. In contrast to the advantages of the asymptotic model, the main disadvantage is that if the total number of observations is very low then the results might be invalid. Nevertheless, despite all the advantages and disadvantages of both simulation strategies, the logistic model should be efficiently modeled to correctly predict the outcome.

The Weibull competing risk model plus the posterior predictive model of the events are combined in one of the Metropolis Hasting algorithms. Each of the simulations of those quantities such as coefficients, predicted values, shape, and scale parameters are kept to be aggregated in the end. The Metropolis Hasting is used because only the non-normalized aspect of the posterior distribution is needed or in other words, there is no need to utilize the full posterior distribution. Another advantage of the Metropolis Hasting is that the simulations are done by a generator and not the actual posterior. The generator in our problem follows univariate or multivariate normal distribution with a mean equal to the previous value and a fixed covariance matrix. This type of Metropolis Hasting is called Random Walk. This Random Walk Metropolis hasting has two advantages first the mean of the generator is known and second, the covariance matrix is tuned according to the acceptance rate. When the dimension of the parameter is one acceptance rate of 0.5 is valid and if the dimension is big enough an acceptance rate of around 0.25 is efficient. The convergence defines the iteration number which is tuned appropriately for the stationary and mixing properties to be valid. The only aspect that is somehow arbitrary is the parameters of the priors'

distribution. The prior distribution is defined either from previous knowledge or by sensitivity analysis or arbitrary or non-informed options. In this case, the prior distributions have a mean equal to the null hypothesis and covariance matrix or variance big enough. Before entering the Metropolis Hasting algorithm, itself, the conditional posterior distribution of the desirable parameters is given. The coefficient posterior distribution of the logistic model which produces the probabilities of the predicted posterior Bernoulli distribution, basically is formed as

$$p(\theta|y,z,a_{-\theta}) \propto \exp\left(-\frac{(\theta-\hat{\theta})^T \widehat{I(\theta)} (\theta-\hat{\theta})}{2}\right)$$

Where $\hat{\theta}$ is the likelihood estimation of $\prod_{i=1}^{n} \frac{\exp(y_i \theta' z_i)}{1+\exp(\theta' z_i)}$ and the $\widehat{I(\theta)}$ is the estimated information matrix using the complete dataset. Both those quantities are found from base function of the statistical program and the values of θ are easily simulated from the normal multivariate distribution. Also, $a = \{y_{miss}, \theta, \beta_1, \beta_2, \lambda_1, \gamma_1, \lambda_2, \gamma_2\}$ is the full set of the total parameters and $a_{-\theta}$ is the a expect θ. For the coefficients of the first and second event type model, the conditional posterior distributions are

$$p(\beta_1|x,t,\delta_1,a_{-\beta_1}) \propto$$

$$\prod_{i=1}^{n} [\exp(\beta_1' x_i)]^{d_{1i}} \exp\left(-\exp(\beta_1' x_i)\, \lambda_1 {t_i}^{\gamma_1}\right) \exp\left(-\frac{(\beta_1-\mu_{01})^T {\Sigma_{01}}^{-1} (\beta_1-\mu_{01})}{2}\right)$$

And

$$p(\beta_2|x,t,\delta_2,a_{-\beta_2}) \propto$$

$$\prod_{i=1}^{n} [\exp(\beta_2' x_i)]^{d_{2i}} \exp\left(-\exp(\beta_2' x_i)\, \lambda_2 {t_i}^{\gamma_2}\right) \exp\left(-\frac{(\beta_2-\mu_{02})^T {\Sigma_{02}}^{-1} (\beta_2-\mu_{02})}{2}\right)$$

The conditional posterior distributions for the scale and shape parameters of both events models; namely, λ_1, γ_1, λ_2, γ_2 are

$$p(\lambda_1|x,t,\delta_1,a_{-\lambda_1}) \propto \prod_{i=1}^{n} [\lambda_1]^{d_{1i}} \exp\left(-\exp(\beta_1' x_i)\, \lambda_1 {t_i}^{\gamma_1}\right) g(\lambda_1, a_{\lambda_1}, c_{\lambda_1})$$

$$p(\gamma_1|x,t,\delta_1,a_{-\gamma_1}) \propto \prod_{i=1}^{n} [\gamma_1 {t_i}^{\gamma_1-1}]^{d_{1i}} \exp\left(-\exp(\beta_1' x_i)\, \lambda_1 {t_i}^{\gamma_1}\right) g(\gamma_1, a_{\gamma_1}, c_{\gamma_1})$$

$$p(\lambda_2|x,t,\delta_2,a_{-\lambda_2}) \propto \prod_{i=1}^{n} [\lambda_2]^{d_{2i}} \exp\left(-\exp(\beta_2' x_i)\, \lambda_2 {t_i}^{\gamma_2}\right) g(\lambda_2, a_{\lambda_2}, c_{\lambda_2})$$

$$p(\gamma_2|x,t,\delta_2,a_{-\gamma_2}) \propto \prod_{i=1}^{n}[\gamma_2 {t_i}^{\gamma_2-1}]^{d_{2i}} \exp\left(-\exp(\beta_2' x_i)\,\lambda_2 {t_i}^{\gamma_2}\right) g(\gamma_2, a_{\gamma_2}, c_{\gamma_2})$$

All likelihoods start from the first observation, and they are ending to the nth observation. This is because the first simulated values are the missing events which are simulated by the Bernoulli distribution. So, the simulation of the previous parameters is conducted after the imputation of the dataset. Last but not least the generator of the candidate simulations is normal distribution either univariate or multivariate with a mean equal to the previous simulated values and a fixed variance or covariance matrix respectively (it is tuned by the acceptance probability). The candidate simulations of the univariate normal distribution are either positive or negative; as a result, sometimes the candidate simulations of the scale and shape parameters are negative which is not valid. Therefore, when a negative value is simulated in a specific iteration the probability of the acceptance in this specific iteration is basically zero.

In conclusion, first, simulate the missing event types and impute those values in the dataset. After that update the parameters of the predictive posterior distribution using the complete imputed dataset. The simulation is conducted through the conditional posterior distribution. Utilizing the complete dataset update the coefficients of the first and the second event model. In the end, the univariate parameters of the two Weibull models are updated. The simulations are conducted via the generator distribution which is univariate or multivariate normal distribution with mean equal to the previous values and fixed covariance matrix. The acceptance probability for each simulation in each iteration is calculated which is basically the minimum between one and the multiplication of the condition posterior density of the candidate simulation and the density of the generator on the prior simulation divided by the multiplication of the conditional posterior density on the previous simulation and the density of the generator on the candite or possible value. For instance, the probability of acceptance in the $i+1$ iteration of the parameter θ in Metropolis Hasting Random Walk is

$$p(\theta^{pos}, \theta^i) = \min\{1, \frac{p(\theta^{pos}|\alpha^i_{-\theta})}{p(\theta^i|\alpha^i_{-\theta})}\}$$

Where $\alpha^i_{-\theta}$ is the latest set of the simulation without the θ^i, $p(\theta|\alpha^i_{-\theta})$ is the posterior probability of a given parameter θ and the θ^{pos} is the generated – possible new simulated value if it is accepted. Finally, after the implementation of numerous simulations, the convergence criteria are checked. If the convergence criteria are not satisfied, then the number of iterations is increased. If the convergence criteria are satisfied, then the simulations are aggregated for the final – aggregated model to be described. After the aggregation of the simulated coefficients, an analyst should proceed to a model evaluation. Therefore, for each model according to the number of event types residuals like Cox-Schnell, martingale, and deviance residuals are calculated.

Apart from the mean and the standard deviation of each parameter the central interval (CI), the probability of direction(PD), and the region of practical equivalence (ROPE) are given [26,27].The 95% central interval has been thoroughly described in the second chapter, it is the 2.5% and 97.5% quantiles of the simulations. Its interpretation is that the probability of the parameter belonging to this interval is 0.95. Nevertheless, the probability of direction and the region of practical equivalence have not been efficiently described. The probability of direction [26,27] measures the effect existence of a specific estimated parameter. Particularly it measures how positive or negative the effect of a coefficient and it ranges from 0.5 to 1. Also, it is calculated from the simulations, it is robust to the scale of both outcome and the covariates, and it is somehow empirically related to the frequentist's prominent p-value. It is highly correlated to the p-value and there is an empirical relationship that connects those two. In a one-sided hypothesis the p-value is empirically almost equal to one minus the probability of direction and the two-sided p-value is empirically equal to two multiplied by one minus the probability of direction. The probability of direction is the proportion of the same simulation's sign with the median sign. In other words, it is measured the quantity of the simulations having the same sign as the median. Finally, it measured the existence of the relation, positive or negative, and not the power of the relation. For instance, coefficients might have a big probability of direction but low power which is the low distance between the coefficient and zero. The region of practical equivalence [26,27] has a center equal to the null hypothesis. This equivalence test measures the proportion of standardized full(100%) CI being intersected by the ROPE. The ROPE is suggested [30] being $[-0.1,0.1]$ supposing the parameter is standardized (or all the simulations are divided by the standard deviation). It measures the significance or the power of a parameter estimation. A small percentage of 100% central interval being part of the rope means that the significance of the estimated coefficient is big. In other words, ROPE is the interval of no effect and if a big percentage of the standardized range is inside the ROPE then the null hypothesis is more plausible to be accepted. Nevertheless, this equivalence test is very objective and there are many different ROPE options. Both the probability of direction and the region of practical equivalence indicate efficient evidence that there is a relation between the outcome and the covariate. Nevertheless, the Bays factor is the most traditional option but in this problem, the calculation of the Bayes factor is not done by ordinary methods like Monte Carlo and that is because the dataset is incomplete. Therefore, the computation of Bayes factors in a dataset with missing event type is a big and complex story which in this scenario will not be analyzed any further.

All in all, the Bayesian Weibull parametric competing risk model with missing cause of failure is a combination of two methodologies the first methodology is about a Bayesian way to handle missing values [17] and the second one is about the application of the Bayesian Weibull parametric competing risk model [29]. The basic algorithm for each iteration is :

- Step 1: Impute the missing cause of failure using a predictive model. Specifically, the predictive model is a general linear model with Y equal to one for the first cause of failure and zero for the second one. The covariates are the natural

logarithm of the time to the event, the Gender, Age, and CD4. The coefficients of the predictive model are estimated in the second step. Therefore, in the first iteration, the initial values of the model are estimated before the iterations begin, and they are estimated via the maximum likelihood method. The predictions are derived from the predictive probability using the ordinary Monte Carlo method.

-**Step 2:** Update the parameters of the predictive model. This is done by first estimating the coefficients of the predictive model using all the datasets. Those coefficients are estimated using the maximum likelihood method. Finally, update those coefficients by simulating them from a multivariate normal distribution with a mean equal to the already estimated coefficients and covariance matrix the inverse information matrix which is asymptotically equal to the estimated covariance matrix of the coefficients.

-**Step 3:** Simulate first the coefficients, the scale, and then the shape parameters of the first event and then simulates like the first event the coefficients, scale, and shape parameters of the second event using the Random Walk Metropolis Hasting algorithm. The coefficients from both models are simulated from a multivariate normal distribution with a mean equal to the previous accepted value and with a covariance matrix which is tuned via the mean acceptance probability. The scale and shape parameters of both events are generated from a univariate normal distribution with a mean equal to the previously accepted simulations and variance tuned from the mean acceptance probability. The covariance matrix and the variance of the generators are calculated by trial and error with the aim of the mean acceptance probability being approximately equal to 0.4 for the multivariate coefficients and 0.5 for the univariate parameters (scale–shape). In every iteration, the generated values either are accepted or not. If they are not accepted, the previous values instead of the current ones are the accepted ones.

In the end, there are numerous simulations. From all those, I burn some initial simulations, and then because of the high autocorrelation I choose one simulation every 25 ones. From the final simulations, the desirable coefficients are the mean of the corresponding simulations.

4.1.3 Cox competing risks with missing event types

This method is based on a sophisticated computationally maximum pseudo-partial-likelihood estimation method for the semiparametric (Cox) proportional cause-specific hazard model [16]. In other words, a Cox competing risk model on a weighted complete dataset is implemented. Virtually, weights and event indicators are assigned to the missing event types resulting in a weighted full dataset. The censored times are randomly right censored, and the missing cause of failure is under the missing at-random assumption. The maximum pseudo-partial-likelihood(MPPL) estimation approach is used to calculate the coefficients when the likelihood is not the ordinary one. For example, a weighted likelihood is calculated by the MPPL method. This method utilizes the same event probability function or in this particular scenario a logistic regression model as the corresponding Bayesian approach. Particularly the same model as the event

imputation method in the Bayesian approach is used. To begin when the event types are two, a logistic regression model is applied to the observations with a specific event type (the first or the second cause of failure). Then after the estimation of one complex logistic model with interactions, splines, fractional polynomials, and more, the missing observations are duplicated. In the first bunch, the missing event indicator is assigned as the first event with weights equal to the probability of this observation having the first event type. In the second bunch, the second event indicator is assigned, and their weights are equal to the probability of having this event. This logistic model is calculated at the start only by the observed noncensored observations basically only of those who have experienced an event. Therefore, after the dataset preparation, two separate usual Cox regression models are applied with the aim of estimating the coefficients. The standard error of the coefficients is estimated by bootstrap methods. Especially, the initial given dataset is resampled with replacement and then the same analysis is conducted again and again. Then after the recalculation of the coefficients, the standard deviation of the recalculated coefficients is the actual standard error of the initial coefficient. The covariance matrix is estimated by the sample covariance matrix which is calculated by the bootstrapped sample. Finally, the typical Cox model evaluation is conducted which evolves the calculation of Cox-Schnell, martingale, and deviance residuals. In addition, the Schoenfeld residuals are calculated to check the proportional hazard assumption.

Let assume that there are n observations, 2 cause of failures, the d_{ki} indicator which is basically equal to 1 if the ith individual has experience the kth event and 0 if it is not $i = 1, 2,3, \dots, n$ and $k = 1, 2$, the time t_i until the experienced event or the end of the survey or a lost of follow up, a set of predictors $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ for each individual where p is the number of variables, an auxiliary set of covariates $z_i = (z_{i1}, z_{i2}, \dots, z_{im})$ which are function of x_i and the time to event and there n_1 missing events. The missing event types are under missing at random assumption and the censoring is independent right. To begin with the logistic model is calculated with only a part of uncensored $n - n_1$ observations. This model includes the predictors of the primary model, some auxiliaries covariates, and the survival time. The outcome of this model is an indicator that is equal to one if the event type is the first one and is equal to zero if the cause of failure is the second one. The weights of all observations except the missing ones are equal to one. The rest unobserved data are imputed with the first event indicator with the corresponding predicted probability as weight. Finally, the same imputed part of the dataset is copied, the first event type is replaced with the second, and the weights are replaced with the one minus the previous weights. The copied imputed dataset is aggregated to the first semi-imputed dataset. Finally, the complete aggregated dataset proceeded to the final typical cause-specific competing risk analysis.

This method in contrast to the multiple imputation method proposed by [9,15] has shown better empirical results [16]. Also, the MPPLE method is computationally lighter and easier than the multiple imputation method because virtually it just assigns some weights to a specific observation and after that, the typical Cox model is implemented. In addition, both methods need this logistic regression model. In the MPPLE method only at the start, a logistic regression

model is estimated in order to predict the probabilities in opposition to the Bayesian method which the predictive logistic regression model is re-estimated in every iteration in order to provide the necessary probabilities. Also, if the dataset is too big the multiple imputation is a cumbersome and tedious method to utilize. Last but not least, the purpose of this supplementary method is to compare the results, therefore it is necessary that the second method is easy and straightforward to implement and that is the final reason why MPPLE has been chosen.

4.1.4 Conclusion

To sum up, the purpose of this thesis is to implement the Bayesian parametric competing risk in missing event types. The second purpose is to compare the results of this method to another method like the method that is based on the maximum pseudo-partial likelihood. Both methods utilize a logistic regression method; as a result, this general linear model must be carefully modeled and checked. So, the introduction of splines, infections, fraction polynomials, high polynomial powers, and more is necessary. In addition, various model-checking methods are applied to evaluate the competence of the logistic model. The most ordinal methods are the Chi-square test, the deviance test, and Hosmer Lemeshow test. That goodness of fit test assures the analyst that the applied logistic model is a good and valuable fit. The high effectiveness of the logistic regression model is a significant issue because both methods utilize this model. Finally, the analysis is implemented in R version 4.3.0 and the algorithms for both models are written in R by me except the Cox and logistic regression. For the Cox competing risk model with missing failure, I have copied the implementation algorithm from the appendix of [16]. All the code from R is given in the Appendix.

4.2 The Data

4.2.1 Real Data

The analysis is supposed to be conducted with data from the EA-IeDEA HIV study (link for the site https://www.iedea.org/regions/east-africa/). EA-IeDEA stands for the East African International Epidemiologic Databases to Evaluate AIDS and it is basically a huge endeavor to evaluate and eventually tackle the HIV disease in the East African Region. However because of bureaucratic reasons, the data were not immediately available, and a long procedure was needed in order to acquire a dataset from the EA-IeDEA HIV study. Therefore, the main thesis goal is changed, and it becomes a simulation study that is basically trying to compare the two methods first the Bayesian Weibull competing risk with missing cause of failure method with the maximum pseudo-partial-likelihood estimation method (MPPLE) which is thoroughly described in [16]. This MPPLE method imputes probabilities as weights to the observations with missing event types and then the standard weighted Cox regression is implemented. My master thesis simulation study tries to imitate the EA-IeDEA data that appeared to [16]. This data appeared to [16] have two event types the first one is death and the second one is disengagement from HIV care. Also, a large proportion of patients were still in care when the study was finished and therefore those patients were characterized as censored observations. Apart from that, less than half patients were missing and some of them were reached by healthcare professionals. In other words, some patients missed the schedule leading to a huge proportion of missingness. Because, the actual observed events were small, relative to the censored and missing observation, healthcare professionals tried to reach those missing patients. From the outreached sample of the initial missing patients, scientists found out that most of them were disengaged and 22% of them died leading to significant reporting death problems. Except for the time to the event or missingness or censoring, the gender, the age of patients at the ART initiation in years, and the number of CD4 cells at ART initiation in cells/μl were counted. In this EA-IeDEA study sample the number of patients were 6657 HIV-infected patients on ART. 3382 of them were in care when the study was finished and the rest of them 3275 were either missing or dead. Only 346 patients from 3275 died and therefore they were reported in the clinic as dead ones. The rest 2929 patients were missing and after the outreach only 448 were successfully reached by the hospital workers. 349 of them were disengaged from HIV care and 99 died. Eventually, this is leading to 2481 missing observations and the actual percentage of missingness given that the observation is not censored is 75.8% which is significantly large. The percentage of censored observations and the observed cause of failure being death is 50.8% and 56% respectively. The 56% percentage indicates that 56 patients out of 100 who have experienced one of two events have died.

Also , the descriptive statistics along with the results [16] have a meaningful role in the simulation study because the information about the distributions for the gender, age and CD4 cells count and the hazard functions of two events is taken from the descriptive statistics table (table 3 in [16]) and hazard ratio table (table 4 in [16]). As a consequence, they are represented in this thesis.

Table 1: Descriptive statistics for the EA-IeDEA study sample [16].

	Cause of failure			
	In Care (N=3382) n(%)	Disengagement (N=349) n(%)	Death (N=445) n(%)	Missing (N=2481) n(%)
Gender Female	2300(68.0)	210(60.2)	254(57.1)	1665(67.1)
Male	1082(32.0)	139(39.8)	191(42.9)	816(32.9)
	Median(IQR)	Median(IQR)	Median(IQR)	Median(IQR)
Age*	37.9(31.8,45.4)	35.5(29.7,41.9)	37.3(31.3,46.0)	35.4(29.9,42.7)
CD4**	174(91,258)	145 (69, 222)	88 (39, 180)	155 (71, 214)

*At Art initiation in years, **At Art initiation in cells/μl

From the Table 1, I try to imitate the distribution of Gender, Age and CD4. Also the hazard ratio and its standard deviation are needed for each variable (Gender, Age and CD4). They are needed for the modelling of the two hazard functions and the covariance matrix of prior distribution respectively. This Table 2 which is represented beneath is part of Table 4 [3], particularly it is only the proposed MPPLE part.

Table 2:Data analysis with the MPPLE method EA-IeDEA study sample [3].

Covariate	$\exp(\beta_n)$	95%CI	p-value
Disengagement from care			
Sex(male=1, female=0)	1.15	(1.02,1.31)	0.022
Age(10 years)	0.75	(0.7,0.8)	<0.001
CD4(100 cells/μ1)	1.03	(1,1.06)	0.094
Death while in care			
Sex(male=1, female=0)	1.24	(0.96,1.59)	0.094
Age(10 years)	1.10	(0.97,1.25)	0.153
CD4(100 cells/μ1)	0.76	(0.63,0.91)	0.003

To begin with, the standard error of the hazard ratios which is used to construct 95%CI of Table 2 is calculated with the bootstrap method [16]. Virtually, the initial sample is resampled with replacement and for each resampled dataset, the coefficients for each event are calculated. Then the standard error of the hazard ratios is the standard deviation of the "resampled" coefficients. In other words, the resampling of the initial dataset is conducted for several rounds, and for each round, a set of coefficients (2 causes of failure) is calculated, then the standard deviation of the coefficients for each covariate is the standard error for each respectively. So, the standard error of the hazard ratio is used to calculate the 95% confidence intervals. Someone can instantly take the q0.025 and q0.975 of the "resampled" coefficient and find the 95% CI instead of calculating the standard error where qp is the quantile that cumulative function is equal to p. Now, because I want the actual log – hazard ratios and their standard errors, I need to transform the values in Table 2. For example, the log hazard ratio of Sex in Disengagement is just ln(1.15) but the log hazard ratio of Age(10 years) in Disengagement is

$\frac{\ln(0.75)}{10}$. The standard error of log hazard ratio is calculated with pretty much the same reasoning. For example, the standard error of the Age(10 years) in disengagement is mean of $|\frac{\frac{\ln(0.7)}{10}-\frac{\ln(0.75)}{10}}{1.96}|$ and $|\frac{\frac{\ln(0.8)}{10}-\frac{\ln(0.75)}{10}}{1.96}|$. Mean is used because there is a possibility that the 95%CI is either calculated using standard errors or the ordinary bootstrap way (just find the quantiles of 0.025 and 0.975 cumulative probability). So, the standard errors, I calculate from the Table 2, are estimations of the real ones and they are used to define the prior distribution of the covariates. The upcoming Table 3 has the transformed hazard ratios which are used to define the hazard function and their standard error.

Table 3: log -hazard ratios and their estimated standard errors

Disengagement from care	Sex(male=1,female=0)	Age(1 year)	CD4(1 cell/μ1)
log – hazard ratio	0.1397619*	-0.02876821	0.000295588
standard error	0.06383278	0.003406413	0.0001486452
Death while in care	Sex(male=1,female=0)	Age(1 year)	CD4(1 cell/μ1)
log – hazard ratio	0.2151114	0.009531018	-0.002744368
standard error	0.1287133	0.006469458	0.0009380734

*The values are exactly the same as the R software computed and they have the maximum potential decimals because those numbers are essential to calculate the bias in the end.

4.2.2 Data Simulation

The are two problems that arise when someone wants to simulate a competing risks scenario with missing events. First, how to simulate the time and the cause of failure. Second, how to determine the distribution parameters for the distribution of Gender, Age, CD4, shape, scale, censored parameters, the model for the missing values indication, and more.

Initially, the time is simulated using the survival function (or the cumulative function) with Monte Carlo simulation methods like inverse probability simulation. In this method basically, someone simulate a random value from a uniform distribution (0,1) and then they find the requested simulated time which satisfies or solves the equation $S(t) = u$ where u is the random value of the uniform distribution. There are two problems, first what is the actual form of $S(t)$ and can this equation be solved with standard ways? Firstly, there two hazard functions because there are two competing risks [16,31], so the survival function is written as $S(t) = \exp(-\int_0^t h(v)dv)$ where $h(t) = h_1(t) + h_2(t)$, $h_1(t)$ is the hazard function related to the death event and $h_2(t)$ is the disengagement hazard function. Those hazard function are defined according to the scenarios, in this thesis there are three scenarios the first one is a Weibull with shape parameters

equal to one, the second is again Weibull with shape parameters different to 1 and in the thirst scenario time follows the Gompertz distribution. The $S(t) = u$ is solved with numerical algorithms like the Newton -Rapson. The solution is unique because the survival function is monotonous function. For all three scenarios, 1000 rounds of simulations are conducted.

After accumulating the survival time, the cause of failure needs to be simulated. There are two causes of failures death and disengagement, those indicators are simulated using Bernoulli distribution. Death cause of failure is simulated with probability p and the disengagement with probability $1 - p$. The probability p is equal to $\frac{h_1(t)}{h_1(t)+h_2(t)} = \frac{\frac{h_1(t)}{h_2(t)}}{1+\frac{h_1(t)}{h_2(t)}}$ [31]. In other words, both $h_1(t)$ and $h_2(t)$ are found for each individual and the probably $p = \frac{h_1(t)}{h_1(t)+h_2(t)}$ is calculated which is the probability experiencing the first event (death).

Nevertheless, the hazard functions are functions of other independent variables like Gender, Age, and CD4. Therefore, the problem is to assign a distribution to them that best matches the descriptive statistics in Table 1. To begin with, the Gender distribution is a Bernoulli one with the probability of being male equal to the males of the study divided by the total sample population. So, Gender follows Bernoulli(0.335) but Age and CD4 are the trickiest. I assume at the start that Age and CD4 follow a normal distribution with means equal to the weighted medians of Table 1 and with variance, a variance in which the q0.25 and q0.75 are close to the weighted quantiles of Table 1. For example, the weighted mean–median of Age distribution is $\frac{37.9*3382 + 35.5*349+37.3*445+35.4*2481}{6657}$ and corresponding mean of CD4 distribution is $\frac{174*3382+145*349+88*445+155*2481}{6657}$. This procedure is repeated for the q0.25 and q0.75. After acquiring the weighted q0.25 and q0.75 then with the trial-and-error method, I found the variance that has q0.25 almost equal to the weighted q0.25 and then I found a second variance that has q0.75 almost equal to the weighted q0.75. After having the two variances, I calculate the mean of them. Then this variance is the requested one. In the end Age and CD4 are simulated from $N(36.80,9.85^2)$ and $N(159.65,115.4^2)$ respectively but because sometimes Age and CD4 take negative values, I take the absolute value of them. In the end Age and CD4 follow an absolute version of Normal distribution.

Next, there are three more important issues, the first one is related to the censoring, the second one to the assignment of missingness, and the third one with the control of the probability the patients died versus being disengaged. In the EA-IeDEA study sample, the proportion of censoring is 50.8% and it is known that the censoring time follows an exponential distribution with a known rate [16]. Therefore, again with the method of trial and error, a rate for the exponential distribution is searched in order to have a censoring probability equal to 0.508. In practice, the optimal rate is a parameter in which the mean of all censoring probabilities that are found in each simulation round is equal to 0.508. The creation of the missing probability model was hard and somehow I copied the Naïve Bayes logic. Basically, in [16] the model of missing probability is applied to non-censored observations, and the same is implemented here. In practice, the probability model is a function of time to event and the other three covariates

because this type of model satisfies the MAR assumption. The missing probability is a multiplication of two other probabilities, the first one is a completely random choice and the second one is dependent on the descriptive statistics Table 1. The first probability is related to time to event, particularly the observations that have time lower than the mean of all simulated time 0.25 quantiles has 0.4 probability to be missing. The observation which has time to event bigger than the mean of all simulated time 0.75 quantiles has also a 0.4 probability of being missing. As a result, the observations that belong to the middle have a 0.2 probability of being missing. This choice is completely random, and it somehow neglects the observations with big or small time to event time. The second probability is the interested one, it is derived from the multiplication of several densities. Especially, this probability is equal to the missing density divided by the sum of missing and observed density. The missing density is a multiplication of 3 densities, the first one is the probability distribution of gender with the parameter $p = \frac{816}{2481} = 0.33$ Table 1 which is the number of missing males derived by the number of missing patients. The second density is a normal distribution with a mean equal to 155 cells/μ1 and a standard deviation equal to 106.2 cells/μ1 (the standard deviation is calculated in the same way as the previous standard deviation of the Age and CD4 distributions). The third density is a normal distribution with mean and standard deviation equal to 35.4 years and 9.48 years respectively. This missing density counts how much the non-censored observation is fitted to the distribution of missing values. In opposition to the missing density, the observation density counts how much a non-censored observation fits to the distribution of dead or disengaged people. The non-missing Gender distribution is Bernoulli one with probability $p = \frac{139+ 191}{349+445}$, to the numerator are the males who have experienced an event and the denominator is both females and males. The other two densities are normal distributions with a mean equal to 36.51 years and 113.05 cells/μ1 for Age and CD4 respectively. The standard deviation of those two distributions is 10.07 years and 108.4 cells/μ1. The two means of the observed density are the weighted medians of dead and disengaged patients. The standard deviation is calculated with the trial-and-error method, particularly first finding the weighted q0.25 and q0.75 of each covariate. Then, find a standard deviation that the q0.25 is almost the same as the weighted q0.25 and a second standard deviation that the q0.75 is almost the same as the weighted q0.75 and ultimately as standard deviation take the mean of the two previous standard deviations. The probability which is equal to the missing density divided by the sum of observed and missing densities is almost equal to one when the observed density is equal to zero and zero if the missing density is low compared to the observed density. After calculating the multiplication of two probabilities pfinal, which the first one is related to time to the event and the second one to the rest of covariates, one final rate is needed to be calculated with the trial and error method. This rate is multiplied by the pfinal and the result of the two above is used to assign missingness to the observations. For each uncensored observation, a random number that follows a uniform(0,1) distribution is simulated, if this simulated value is smaller than the previous result, the patient is missing. The rate is introduced because the probability of missingness given the subject is uncensored is 0.758, therefore the rate is calculated with the trial-and-error method. The rate is optimized with the purpose of achieving a probability of missingness equal to

0.758. The third important issue is, that the probability of an event being the first one (death) is sometimes bigger than the probability of the second event leading to the totally unbalanced distribution of event type. In order to counter this issue, a final rate is introduced which tries to change the scale parameter of the second event (disengagement) in order to achieve 0.56 probability of the event being death only in the beginning. This rate tries to control the probability of events because it is observed that if the scale of both events are almost random then there is a situation in which all cause of failure is death or disengagement. The only issue, I can't control is the final distribution of the patients who are not censored or missing. In general, the observed dead patients are almost two times more than the observed disengaged ones because the time to the first event is smaller than the time to the second event. As a result, the most censored observation is disengaged patients. One can say that if the patients with the first and second events have different censoring distribution then the problem can easily be solved but the problem is then that the censoring stops being independent right censoring. Independent censoring states that the probability of censoring is random and the same to each patient, or each censored or uncensored patient has the same probability of being censored.

4.2.3 First Scenario

In the first and most plain scenario, the hazard functions for both events are steady as time changes. This is implemented by assigning Weibull distribution to both hazard function with shapes equal to 1. Then, for the first event I take the scale parameter from the [32], it is the scale value of the unadjusted model. So, for the first event the scale of the Weibull parameter is $\exp(-1.2) \approx 0.3$ rounded (in the R program I always use non rounded values 0.3011942). Therefore, If I take the second scale which is in [32], the probability of death is almost 1. So, I need a modified scale for the second event, after some trial-and-error, I find out that $1.95 \times \exp(-1.2) \approx 0.59$ is the desirable scale value for the second event because the probability of being dead is equal to the same sampled probability 0.56. The two hazard functions are

$$h_1(t) = \exp(b_{11} \times Gender + b_{12} \times Age + b_{13} \times CD4) \times scale1$$

$$h_2(t) = \exp(b_{21} \times Gender + b_{22} \times Age + b_{23} \times CD4) \times scale2$$

The first event is death, the second event is disengagement and $b_{11}, b_{12}, b_{13}, b_{21}, b_{22}, b_{23}$ and the log hazard ratios appeared to Table 3. For instance, $b_{11} = 0.2151114$ and $b_{22} = -0.02876821$. The united hazard function is

$$h(t) = h_1(t) + h_2(t)$$

and the simulation of time is implemented through the survival function

$$S(t) = S_1(t) \times S_2(t)$$

Where

$$S_1(t) = \exp(-\exp(b_{11} \times Gender + b_{12} \times Age + b_{13} \times CD4) \times scale1 \times t)$$

$$S_2(t) = \exp(-\exp(b_{21} \times Gender + b_{22} \times Age + b_{23} \times CD4) \times scale2 \times t)$$

So, for every observation a random value for uniform(0,1) distribution is simulated and the simulated time is the solution to the $S(t) = u$ equation. This equation in this scenario has direct solution but I use Newton-Rapson method because I need this method for the rest scenarios. After having the requested times, the probability of cause of failure being dead is $\frac{h_1(t)}{h_1(t)+h_2(t)}$. The cause of failure with known probabilities is easily simulated with the inverse method. In this scenario, If I used the scale for disengagement in [32], almost all causes of failure were death. Therefore, I tweaked scale2 in order to have the mean probability of being dead from all simulations equal to 0.56. Then, I proceed to the second part, the assignment of censoring. I know from the study that 50.8% of observations are censored, also I know from [16] that the censoring time is simulated from an exponential distribution. Then, I simulated the censoring time, and as the final time, I took a minimum of two times the censored and uncensored. The rate of the exponential distribution is estimated with the trial-and-error method and it is equal to 0.547 which causes the mean probability of being censored equal to 0.508. The assignment of missingness is the hard and complicated part. Again, in this situation, there is a desirable rate that is estimated with trial and error with the purpose of achieving a mean missingness of 75.8% on the uncensored patients. The total non-rated missing probability is a multiplication of two probabilities. The first one is a totally random choice, the uncensored observations that have time to event lower than 0.26 have a probability of missingness equal to 0.4. The observations that are between 0.26 and 1.28 have a probability of missingness equal to 0.2 and the rest observations have 0.4. The 0.26 and 1.28 are the mean q0.25 and q0.75 of all simulations. The second probability is a function of Gender, Age, and CD4. It is equal to the density of missingness divided by the sum of observed and missing density. Again, the missing density is equal to $f_{Gender}(gender, 0.33) \times f_{Age}(age, 35.4,9.48) \times f_{CD4}(CD4,155,106.2)$

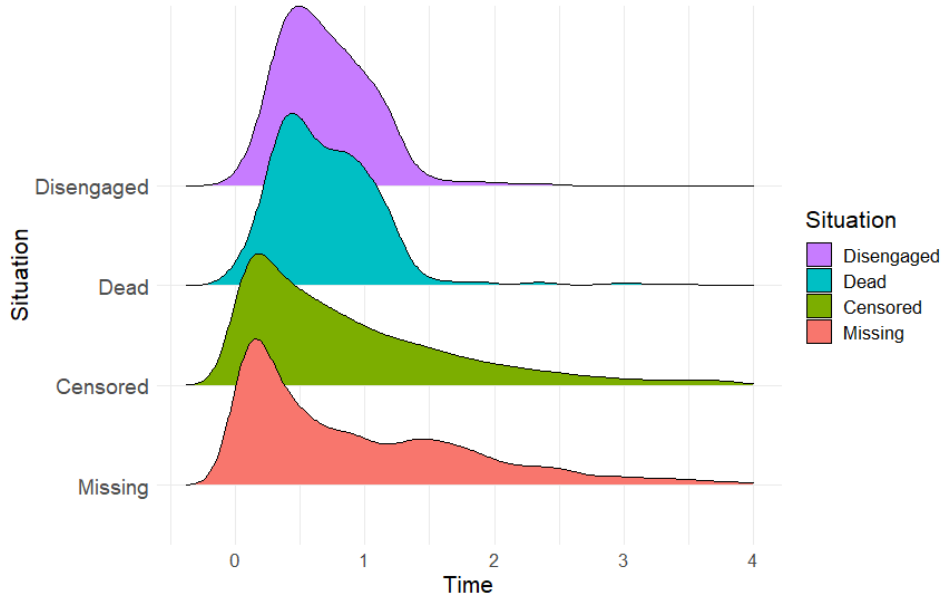
Where f_{Gender} is a Bernoulli distribution with missing probability 0.33, f_{Age} and f_{CD4} are Normal distribution with mean, variance equal to 35.4, 9.48^2 and 155,106.2^2 respectively. Those number are derived from Table 1. The observed density is equal to

$$f_{Gender}(gender, 0.42) \times f_{Age}(age, 36.5,10.08) \times f_{CD4}(113.05,108.4)$$

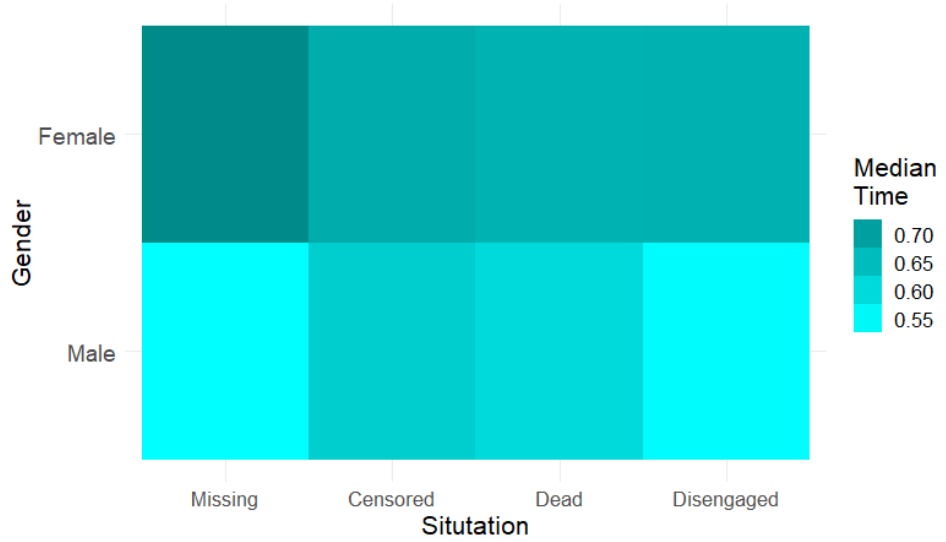
Where f_{Gender} is a Bernoulli distribution with missing probability 0.42, f_{Age} and f_{CD4} are Normal distribution with mean, variance equal to 36.5, 10.08^2 and 113.05,108.4^2 respectively. The way to derive those number is explained in the previous paragraph. Therefore, after calculating the multiplication of two probabilities, I find the appropriate number that if it is multiplied by the previous probabilities , the mean probability of missingness supposing the observation are uncensored is 0.758. This number is equal to 5.329.

Next there are some graphs that are trying to clarify the relations between the variables. Those graphs are from the 250th sample

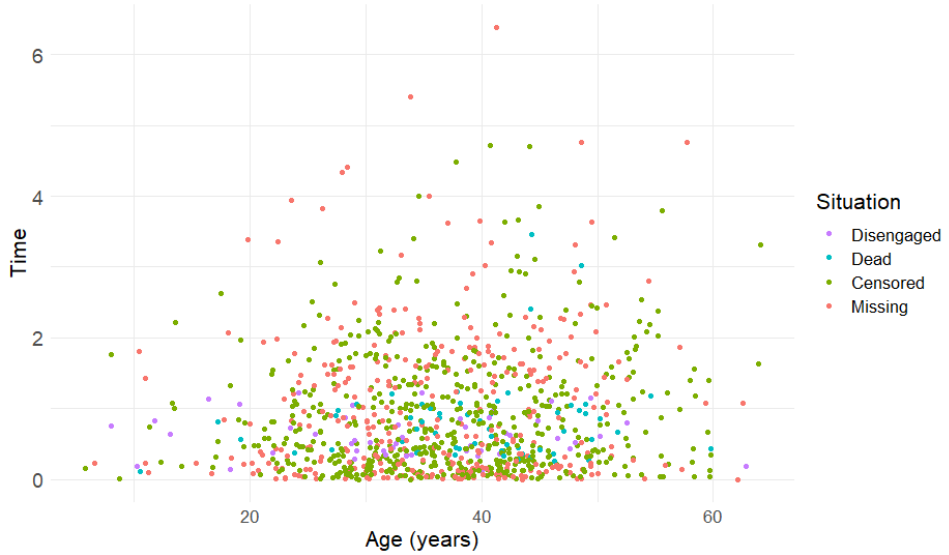
Graph 1: The distribution of time in the four situations (scenario 1)



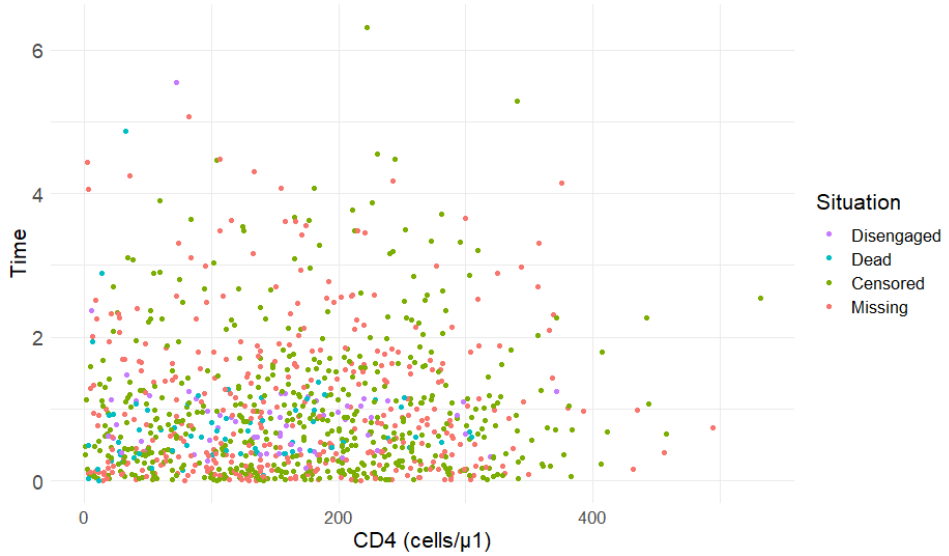
Graph 2: The interaction of time with the gender and with situation (scenario 1)



Graph 3: The relation of time and age (years) (scenario 1)



Graph 4: The relation between the time and CD4 (cells/μ1) (scenario 1)



The observations used for Graph 3 and Graph 4 are 1000 and they are randomly chosen from the 250th dataset. The graphs successfully depict the missing probability model; specifically, most observations with the two desirable events are located between two specific time values 0.26 and 1.28. Also, females have, in general, a bigger median time to either experience the event or be censored or missing. Also, it is observed that the censored times follow an exponential distribution.

Table 4: The descriptive statistics of all simulations.

	Cause of failure			
	In Care (N≈3384)*** n(%)	Disengagement (N≈323) n(%)	Death (N≈470) n(%)	Missing (N≈2480) n(%)
Gender Female	2321(68.6)	186(57.6)	261(55.6)	1664(67.1)
Male	1063(31.4)	137(42.4)	209(44.4)	816(32.9)
Mean:	Median(IQR)	Median(IQR)	Median(IQR)	Median(IQR)
Age*	37.1(30.6,43.7)	35.1(28.2,42.1)	39.2 (32.3,46.2)	36.1(29.5,42.6)
CD4**	169(95,245)	148(69,214)	106(49,178)	163(91, 238)

*At Art initiation in years, **At Art initiation in cells/μl *** The mean number of observations at each simulation.

It is observed that there are various similarities with Table 1. For example, the patients who are still in care have almost the same descriptive statistics as the Table 1. Also, the same applied to the variable Age; namely, the median is bigger at Death and lower at Disengagement and Missing which is similar to Table 1. The percentage of missing female patients is more than the percentage of dead and disengaged ones similar to Table 1. Finally, dead patients have the lowest CD4 and this is similar to the simulated data. In general, the descriptive patterns have been adequately preserved but there are some insignificant number deviances.

Table 5 :Descriptive statistics for the number of patients and time to event

	Cause of failure			
	In Care Median(IQR)	Disengagement Median(IQR)	Death Median(IQR)	Missing Median(IQR)
Patients*	3384(3358,3385)	323(311,333)	470(457,484)	2480(2455,2506)
Time **	0.65(0.27,1.3)	0.63(0.42,0.91)	0.63(0.41,0.92)	0.64(0.2,1.5)

*For each simulation, I keep the number of patients in each category. Those numbers which are indicated are the quantiles of them.
**It is the mean of all medians and IQRs of each simulation.

In practice, there is no actual information about the time to event [16]. Specifically, there are no descriptive statistics about time and the only information given to the paper is about the censoring and the missingness. In addition, it is not known the actual unit of time (days, months, or years) and the implemented Cox models don't give any evidence about the time because there is not a fixed term (proportionally) in opposition to the parametric Weibull model that the scale parameter is basically the fixed term. More information about the simulation code is given in the Appendix.

4.2.4 Second Scenario

The second scenario follows the same methodology as the first one with the difference that the hazard function is not a steady function as time changes. Both the shape parameters of the hazard function are taken from the unadjusted column in [32]. The shape parameter for the first event (death) is equal to 0.584 and the second one is equal to 0.931. The scale for the first event is $exp(-1.2) \approx 0.3$ and for the second event is $1.028 \times exp(-1.2) \approx 0.31$. The scale for the second event is estimated via trial and error and it satisfies the first criteria which is that the mean percentage of the patients being dead is equal to 0.56. The hazard functions are

$$h_1(t) = \exp(b_{11} \times Gender + b_{12} \times Age + b_{13} \times CD4) \times scale1 \times shape1 \times t^{shape1-1}$$

$$h_2(t) = \exp(b_{21} \times Gender + b_{22} \times Age + b_{23} \times CD4) \times scale2 \times shape2 \times t^{shape2-1}$$

The time to event is simulated from the survival function

$$S(t) = S_1(t) \times S_2(t)$$

Where

$$S_1(t) = \exp\left(-\exp(b_{11} \times Gender + b_{12} \times Age + b_{13} \times CD4) \times scale1 \times t^{shape1}\right)$$

$$S_2(t) = \exp\left(-\exp(b_{21} \times Gender + b_{22} \times Age + b_{23} \times CD4) \times scale2 \times t^{shape2}\right)$$

The methodology is exactly the same as scenario one; namely, the time to the event is estimated via the Newton-Rapson method. After, the calculation of the cause of failure being the first event is stimulated using the probability which is equal to $\frac{h_1(t)}{h_1(t)+h_2(t)}$. Then, the second rate is the rate for exponential distribution which the censoring time is simulated. The estimated rate is equal to 0.3435 leading to a 0.508 mean probability of a patient being censored. The 0.25 and 0.75 quantiles of the observed and non-censored time to event are equal to 0.18 and 1.6 respectively. The final rate is equal to 5.36 which causes on average 75.8% of uncensored patients to become missing ones. The rest distributions are the same and the patterns are supposed to be similar.

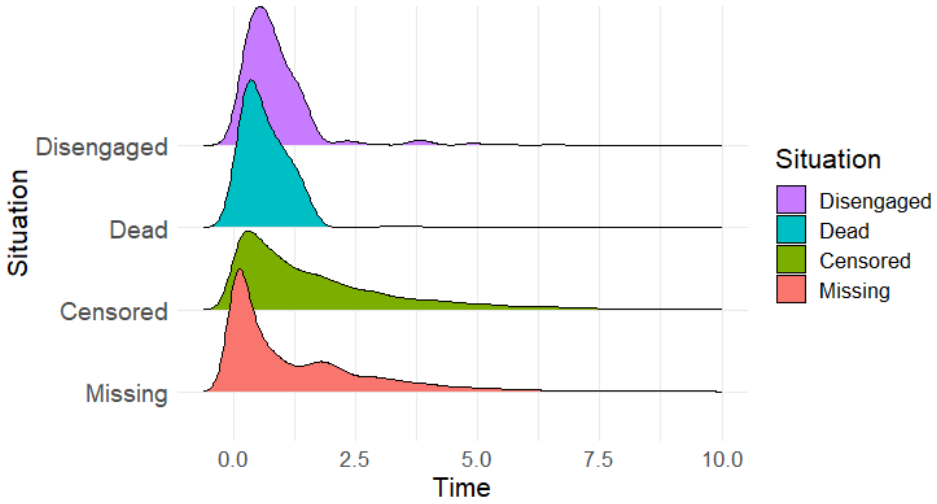
Table 6: The descriptive statistics of all simulations in Scenario 2.

	Cause of failure			
	In Care (N≈3383)*** n(%)	Disengagement (N≈258) n(%)	Death (N≈534) n(%)	Missing (N≈2482) n(%)
Gender Female	2331(68.9)	149(57.8)	295(55.2)	1655(66.7)
Male	1052(31.1)	110(42.2)	239(44.8)	827(33.3)
Mean:	Median(IQR)	Median(IQR)	Median(IQR)	Median(IQR)
Age*	37.0(30.4,43.6)	34.9(28.0,41.8)	39.0(32.1,46.0)	36.2(29.7,42.8)
CD4**	171(97,248)	139(71,215)	106(49,179)	160(89, 235)
Patients****	3383(3355,3411)	258(248,269)	534(520,548)	2482(2455,2510)
Time*****	1.16(0.46,2.42)	0.71(0.4,1.12)	0.56(0.3,0.98)	0.64(0.12,2)

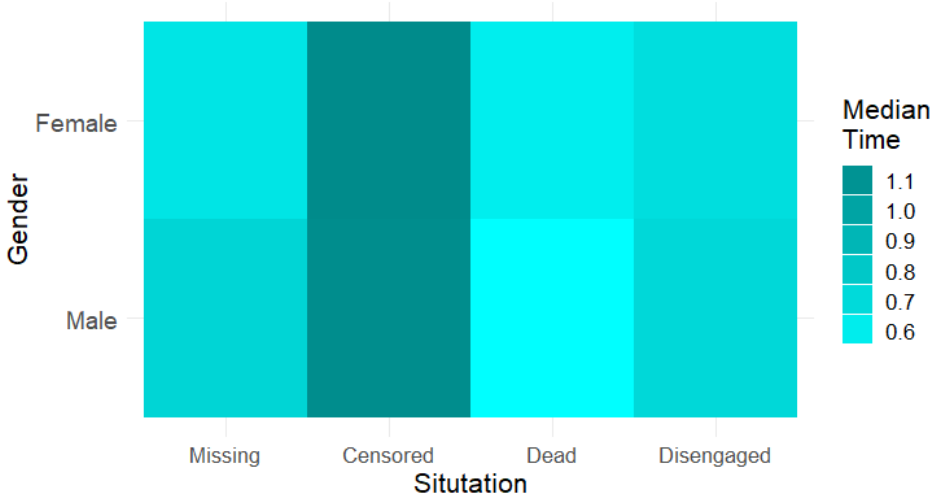
*At Art initiation in years, **At Art initiation in cells/μl, *** The mean number of observations at each simulation, ****For each simulation, I keep the number of patients in each category. Those numbers which are indicated are the quantiles of them. *****It is the mean of all medians and IQRs of each simulation.

It is observed that Table 6 is almost similar to Table 5 and Table 4 which are similar to Table 1. Also, in average 534 patients are dead and 258 are disengaged leading to 2 dead patients for 1 disengaged which is bigger rate than Table 4. The only difference is the distribution of time in the four categories-situations. The median time of patients which are still in care is bigger than the rest of them. This is dissimilar to the Table 5 and therefore some visual information is necessary. This time, I choose the 500th sample of the second scenario to visualize the internal patterns. From graphs, someone can observe that the average median time is pretty much the same in the two genders regardless of the situation. Also, more patients with relatively small time to event are missing compare patients with big times. From graphs, I observe that patients with big times are in the most situations censored.

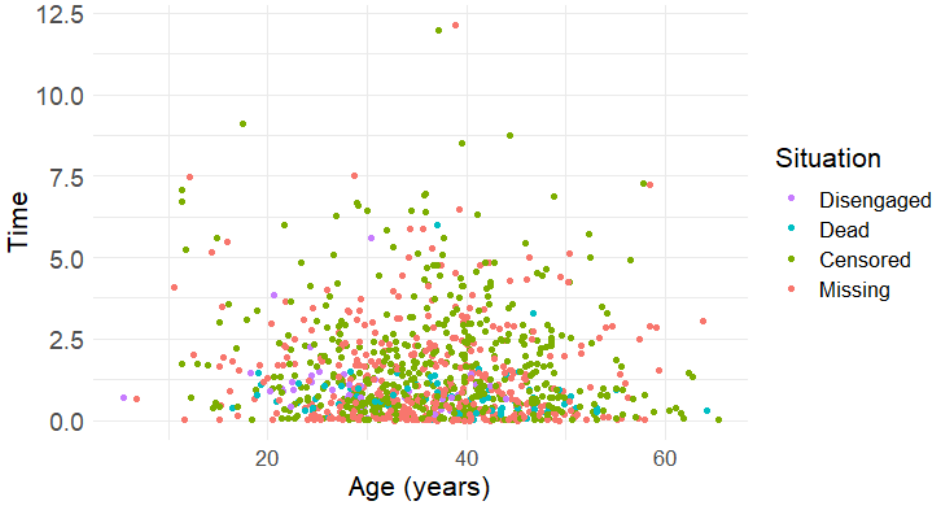
Graph 5: The distribution of time in the four situations (scenario 2).



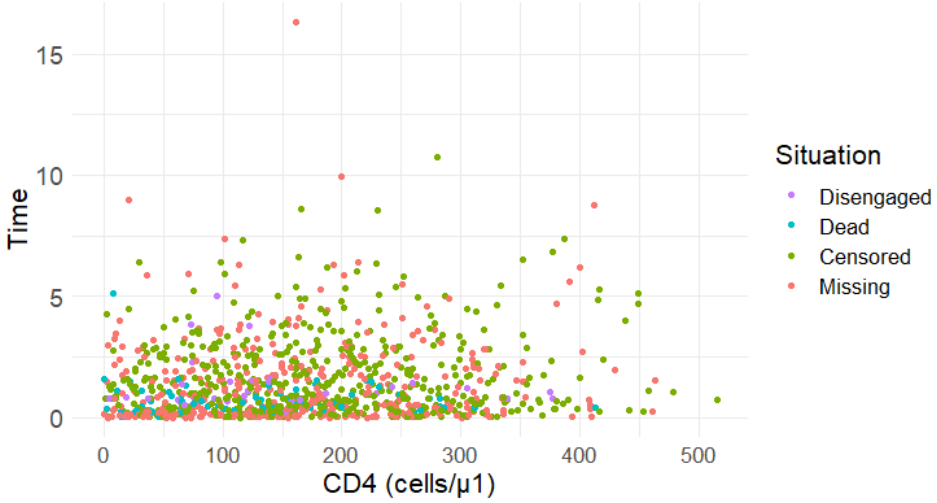
Graph 6: The interaction of time with the gender and with situation (scenario 2)



Graph 7: The relation of time and age (years) (scenario 2)



Graph 8: The relation between the time and CD4 (cells/µ1) (scenario 2)



4.2.5 Third Scenario

The third scenario follows the same methodology as the first and second ones with the difference that time to event follows the Gompertz distribution. Shape parameters of the hazard function are taken from the unadjusted column in [32]. The shape parameter for the first event (death) is equal to 0.584 and the second one is equal to 0.931. The scale for the first event is $0.584 \times exp(-1.2)$ and for the second event is $0.704 \times 0.931 \times exp(-1.2) \approx 0.197$. The scale for the second event is estimated via trial and error and it satisfies the first criteria which is that the mean percentage of the patients being dead is equal to 0.56. The hazard functions are

$$h_1(t) = \exp(b_{11} \times Gender + b_{12} \times Age + b_{13} \times CD4) \times scale1 \times \exp(shape1 \times t)$$

$$h_2(t) = \exp(b_{21} \times Gender + b_{22} \times Age + b_{23} \times CD4) \times scale2 \times \exp(shape2 \times t)$$

The time to event is simulated from the survival function

$$S(t) = S_1(t) \times S_2(t)$$

Where

$$S_1(t) = \exp\left(-\exp(b_{11} \times Gender + b_{12} \times Age + b_{13} \times CD4) \times \frac{scale1}{shape1} \times (\exp(shape1 \times t) - 1)\right)$$

$$S_2(t) = \exp\left(-\exp(b_{21} \times Gender + b_{22} \times Age + b_{23} \times CD4) \times \frac{scale2}{shape2} \times (\exp(shape2 \times t) - 1)\right)$$

The methodology is the same as scenario 1 and 2; namely, the time to the event is estimated via the Newton-Rapson method. After, the calculation of the cause of failure being the first event is stimulated using the probability which is equal to $\frac{h_1(t)}{h_1(t)+h_2(t)}$. Then, the second rate is the rate for exponential distribution which the censoring time is simulated. The estimated rate is equal to 0.517 leading to a 0.508 mean probability of a patient being censored. The 0.25 and 0.75 quantiles of the observed and non-censored time to event are equal to 0.5 and 1.7 respectively. The final rate is equal to 5.306 which causes on average 75.8% of uncensored patients to become missing ones. The rest distributions of the other variables are the same and the patterns are supposed to be similar.

Table 7: The descriptive statistics of all simulations in Scenario 3.

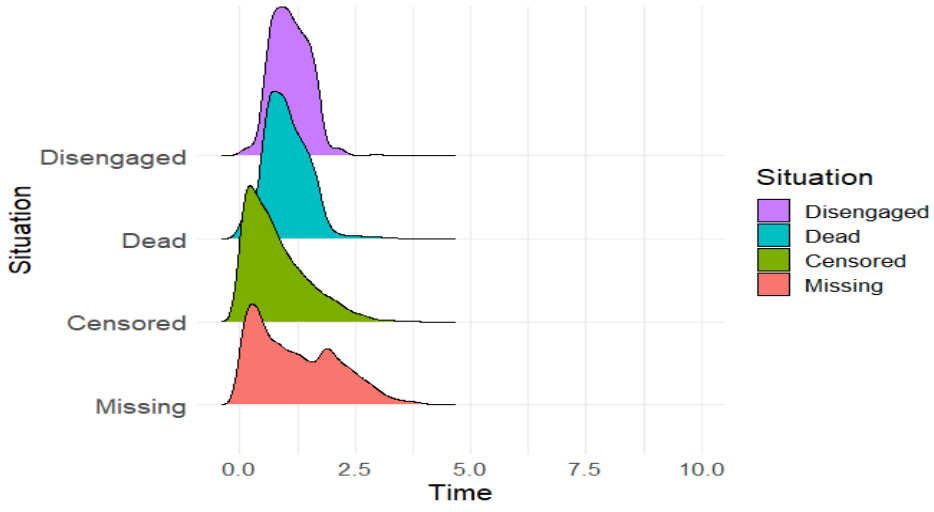
	Cause of failure			
	In Care (N≈3379)*** n(%)	Disengagement (N≈291) n(%)	Death (N≈506) n(%)	Missing (N≈2483) n(%)
Gender Female	2331(68.0)	149(58.0)	295(56.0)	1655(67.7)
Male	1052(32.0)	110(42.0.)	239(44.0)	827(32.3)
Mean:	Median(IQR)	Median(IQR)	Median(IQR)	Median(IQR)
Age*	37.0(30.4,43.6)	35.1(28.2,42.1)	39.1(32.2,46.1)	36.3(29.8,42.8)

CD4**	167(93,244)	140(71,216)	108(50,181)	165(93, 241)
Patients****	3379(3350,3403)	291(278,302)	506(492,522)	2483(2456,2507)
Time*****	1.06(0.39,1.92)	1.08(0.77,1.41)	1(0.7,1.35)	1.06(0.39,1.92)

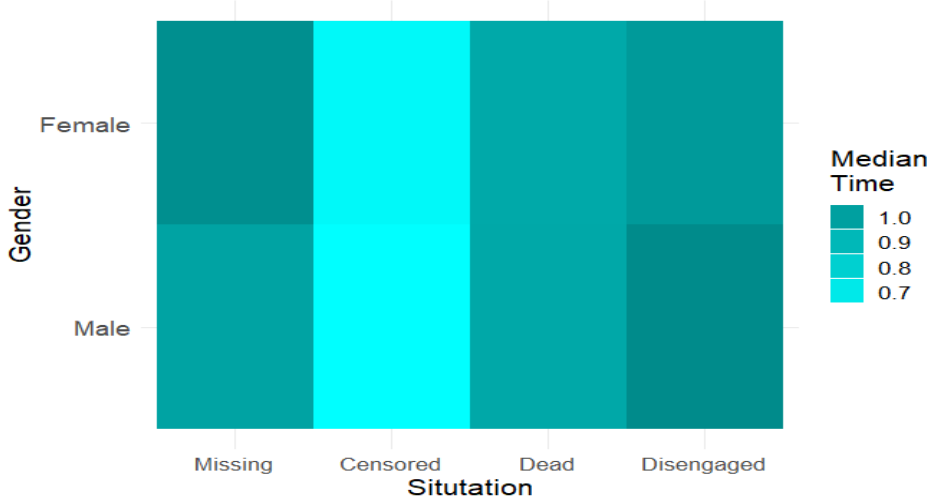
*At Art initiation in years, **At Art initiation in cells/μl, *** The mean number of observations at each simulation, ****For each simulation, I keep the number of patients in each category. Those numbers which are indicated are the quantiles of them. *****It is the mean of all medians and IQRs of each simulation.

It is observed that Table 7 is almost similar to Table 5, Table 4, and Table 6 which are similar to Table 1. Also, on average 506 patients are dead and 291 are disengaged leading to 7 dead patients for 4 disengaged which is a bigger rate than Table 4 but smaller than Table 6. This time, I chose the 750th sample of the third scenario to visualize the internal patterns. From the graphs, someone can observe that the average median time is pretty much the same in the two genders regardless of the situation. Also, more patients with relatively small time to event are missing compared to patients with big times. From the graphs, I observe that patients with big times are in most situations censored. In the following plots, the median time in two genders across all four categories is pretty much the same. It is observed that observations with observed time are between some values (0.25 and 0.75 quantiles). Finally, after probably the 0.75 quantile of the uncensored (before the assignment of missingness) observations, most of the observations are missing.

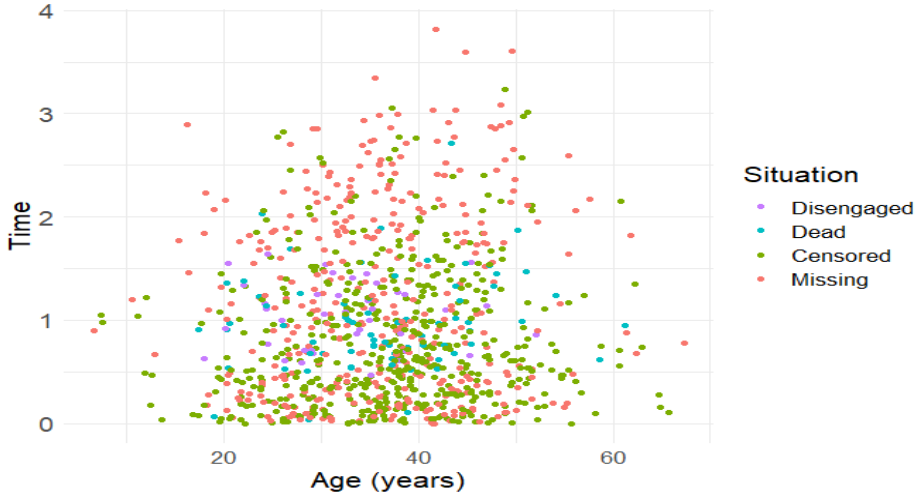
Graph 9: The distribution of time in the four situations (scenario 3).



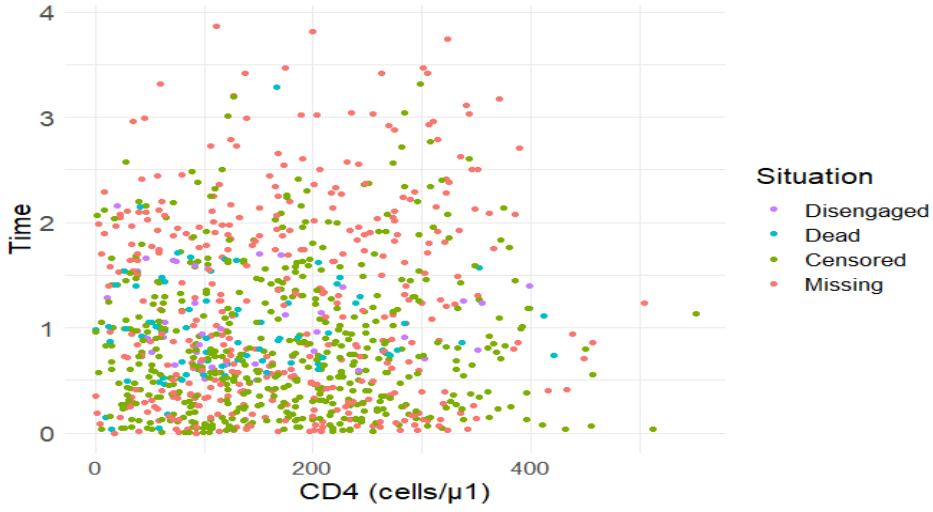
Graph 10: The interaction of time with the gender and with situation (scenario 3)



Graph 11: The relation of time and age (years) (scenario 3)



Graph 12: The relation between the time and CD4 (cells/µ1) (scenario 2)



5 Results

5.1 Bayesian Parametrization

5.1.1 Bayesian Parameters for all three scenarios

For all three scenarios, the prior distribution of the survival coefficients is a multivariate normal distribution with a mean equal to the log hazards ratios of Table 3. I assume that the coefficients are independent with covariance equal to zero and with standard deviation equal to the corresponding standard deviations of Table 3. Therefore, the covariance matrix for each event is a diagonal one with variances equal to the square of the standard deviations. The initial point for the Metropolis-Hastings chains is always equal to the mean of the multivariate normal distribution; namely, the log hazard ratios of Table 3. The generator of the coefficients has a mean equal to the previous simulated vector and covariance matrix which is calculated with respect to the acceptance probability approximately being 0.4. A similar concept is applied to the shape and scale parameters of the Weibull distribution. The prior distribution of the shape and scale parameters has a shape parameter equal to the true value of the parameter and a scale equal to one. The generator of the shape and scale parameters follows a normal distribution with mean equal to the previous value and standard deviation with respect to the probability of acceptance approximately being 0.5. The total number of chains is two and each chain starts from the same initial value and simulates 3000 values, after a 500 burn period, one simulated value per 25 simulated values is chosen. In the end, 100 final simulated values for each chain are chosen leading to 200 final simulated values for each coefficient. Practically, there are 10 desirable parameters for each simulated data. In conclusion, $6000 \times 10 \times 1000 = 60000000$ values are simulated for each scenario, and it took approximately 20 hours to finish one whole simulation for one specific Bayesian scenario. In addition, for every scenario, the coefficients of the predictive model are simulated. The number of predictive coefficients is 5; as a result, there are $6000 \times 5 \times 1000 = 30000000$ extra simulations for every scenario. Therefore, there are three scenarios that lead to 90000000×3 simulations. That is the reason why the two unprocessed chains for each coefficient have only 3000 simulations each and start with the same starting point.

Finally, the missing probability model in both methods plays a crucial role and therefore the true predictive probability model is compared to the final estimated predictive models of both MPPLE and Bayesian methodology. For relative reasons, the model is again described in order for the reader to better understand the model. The Bayesian Weibull Competing Risk model with missing event types has two stages, the first one is the imputation of the missing observations and the update of the missing probability model, and the second stage is the simulation of the coefficients using a complete dataset. In the first stage the missing probability model which is used to impute the missing events, before the loop begins, is estimated using only the observed estimations and then it is re-estimated again with the full complete imputed dataset. After that, the final coefficients of the

probability model are simulated from a multivariate normal distribution with a mean equal to the previous re-estimated coefficients and with a covariance matrix the covariance matrix of the previous re-estimated coefficients. Then in the next round again the missing values are imputed and again the coefficients of the missing probability model are simulated from the multivariate distribution with a mean equal to the re-calculated coefficients which are estimated, using the typical maximum likelihood method, from the last complete imputed dataset. This procedure leads to the final true missing probability model. In other words, the probability model in the end after some burn-in rounds convergences to the true missing probability model. Nevertheless, in this simulation study, it well known from the start the form of the missing probability model and that is the model which describes the probability of the cause of failure being the first one is equal to the hazard function of the first event divided by the sum of the two hazard functions. As a result, a comparison between the estimated and the true predictive model is possible.

Also, for the MPPLE methodology in every dataset, the predictive probability is estimated by the current dataset. The probability model is a general linear model with an outcome equal to one for the first cause of failure and equal to zero for the second one. The covariates for both three scenarios are the natural logarithm of the time to event, the Gender, Age, and CD4. In every dataset, the general linear model is estimated and then its predictive probabilities are the weights of the Cox regression model. Last but not least, from the comparison of the two estimated predictive models with the true one, an analyst can compare which method handles the missing cause of failure. Both final predictive models for every scenario are derived from the coefficient aggregation of the 1000 predictive models.

5.1.2 First Scenario

For the first event (death), the multivariate normal distribution has a covariance matrix equal to $0.00000009\times\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$ leading to a 0.46 mean acceptance probability. For the scale parameter, which is equal to 0.301, the standard deviation of the generator is equal to 0.014 and for the shape parameter, which is equal to one, the standard deviation of the generator is 0.034 resulting in 0.51 and 0.53 mean acceptance probability. The starting points for the scale and shape parameters are 0.301 and 1 respectively. For the second event (disengagement), the multivariate normal distribution has a covariance matrix equal to $0.00000007\times\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$ leading to 0.41 mean acceptance probability. For the scale parameter, which is equal to 0.587, the standard deviation of the generator is equal to 0.034 and for the shape parameter, which is equal to one, the standard deviation of the generator is 0.0395 resulting in 0.49 and 0.52 mean acceptance probability. The starting points for the scale and shape parameters are 0.587 and 1 respectively.

5.1.3 Second Scenario

For the first event (death), the multivariate normal distribution has a covariance matrix equal to $0.0000001 \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ leading to a 0.42 mean acceptance probability. For the scale parameter, which is equal to 0.301, the standard deviation of the generator is equal to 0.013 and for the shape parameter, which is equal to 0.584, the standard deviation of the generator is 0.019 resulting in 0.51 and 0.53 mean acceptance probability. The starting points for the scale and shape parameters are 0.301 and 0.584 respectively. For the second event (disengagement), the multivariate normal distribution has a covariance matrix equal to $0.00000008 \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ leading to 0.41 mean acceptance probability. For the scale parameter, which is equal to 0.31, the standard deviation of the generator is equal to 0.018 and for the shape parameter, which is equal to 0.931, the standard deviation of the generator is 0.038 resulting in 0.52 and 0.51 mean acceptance probability. The starting points for the scale and shape parameters are 0.31 and 0.931 respectively.

5.1.4 Third Scenario

For the first event (death), the multivariate normal distribution has a covariance matrix equal to $0.0000001 \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ leading to a 0.43 mean acceptance probability. For the Gompertz scale parameter, which is equal to 0.176, the standard deviation of the generator is equal to 0.01 and for the Gompertz shape parameter, which is equal to 0.584, the standard deviation of the generator is 0.047 resulting in 0.560 and 0.499 mean acceptance probability. The starting points for the scale and shape parameters are 0.176 and 0.584 respectively. For the second event (disengagement), the multivariate normal distribution has a covariance matrix equal to $0.00000008 \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ leading to 0.404 mean acceptance probability. For the Gompertz scale parameter, which is equal to 0.197, the standard deviation of the generator is equal to 0.022 and for the Gompertz shape parameter, which is equal to 0.931, the standard deviation of the generator is 0.06 resulting in 0.483 and 0.527 mean acceptance probability. The starting points for the scale and shape parameters are 0.197 and 0.931 respectively.

5.2 Results and Simulation Metrics

5.2.1 Evaluation Metrics

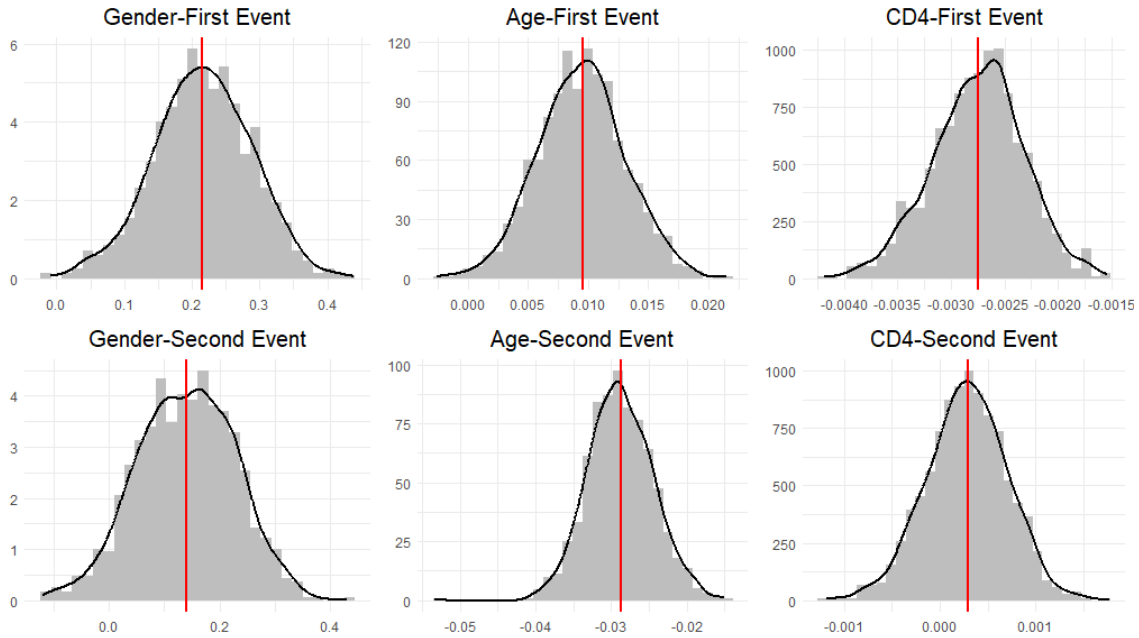
The bias and its Monte Carlo standard error estimation, Empirical Standard Error, Mean Square Error, Average Model Standard error, Relative error of 100%, the 0.025 and 0.975 quantiles and finally a 95% confidence interval of the mean [16,33] are calculated and indicated in the following tables. Virtually, there are 6 tables in which they compare the two methods the standard one MPPLE method and the Bayesian Method. The bias estimate and its Monte Carlo SE of estimate are equal to $\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - \theta$ and $\sqrt{\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2}$ where n are the total number of simulations, $\hat{\theta}_i$ is the estimated coefficient of the i^{th} round of simulations and $\bar{\theta}$ is the $\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$. The Empirical Standard Error is the typical standard deviation of the simulated coefficients which is equal to $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2}$. The Mean Square Error is equal to $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta\right)^2$ which is basically a squared version of bias estimate. The Average Model Standard Error is a square root of mean variance of simulated coefficients which is equal to $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{Var(\hat{\theta}_i)}}$. There is a relation that connects the Empirical Standard Error and the Average Model Standard Error which is $E(AMSE^2) = ESE^2$ [33].The Relative Error $\times$ 100% is equal to $\frac{\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - \theta\right|}{|\theta|}100\%$. The 95% quantile interval of the simulated values that is created from the 0.025 and 0.975 quantiles of all simulated coefficients. Finally, the 95% confidence interval of the mean. The bias estimate, the relative error, and the MSE are a way to check how similar are the mean of the simulated coefficients with the real ones. The Empirical Standard Error is supposed to be equal to the Average Model Standard Error to have an adequate convergence. Finally, the quantile interval shows the variability of the simulated value, and the confidence interval explains the variability of the mean.

5.2.2 Actual Results

For each coefficient of the two events, a matrix is represented which has the previous simulation metrics for both methods and all 3 scenarios. Also, several histograms of the simulated values are represented one for each coefficient and one for the Weibull parameters. The R hat values that describe the convergence of the Bayesian method are given in the Appendix. The rounding has been conducted in four, five, or six decimals. It is observed from Graphs 13-21 that the final chosen simulations follow a normal distribution. From Table 8, I observe that the Bias, MSE, and relative absolute error are bigger in the MPPLE method than the Bayesian one. Also, the target value (the coefficient) belongs to all 95% confidence and quantile intervals. Also, the ESE and AMSE in Table 8 are almost the same in the MPPLE method in opposition to the Bayesian one but the ESE and AMSE are significantly bigger in the MPPLE method than the Bayesian one. In Table 9, the Age coefficient of the first belongs to all 95% confidence intervals except from the

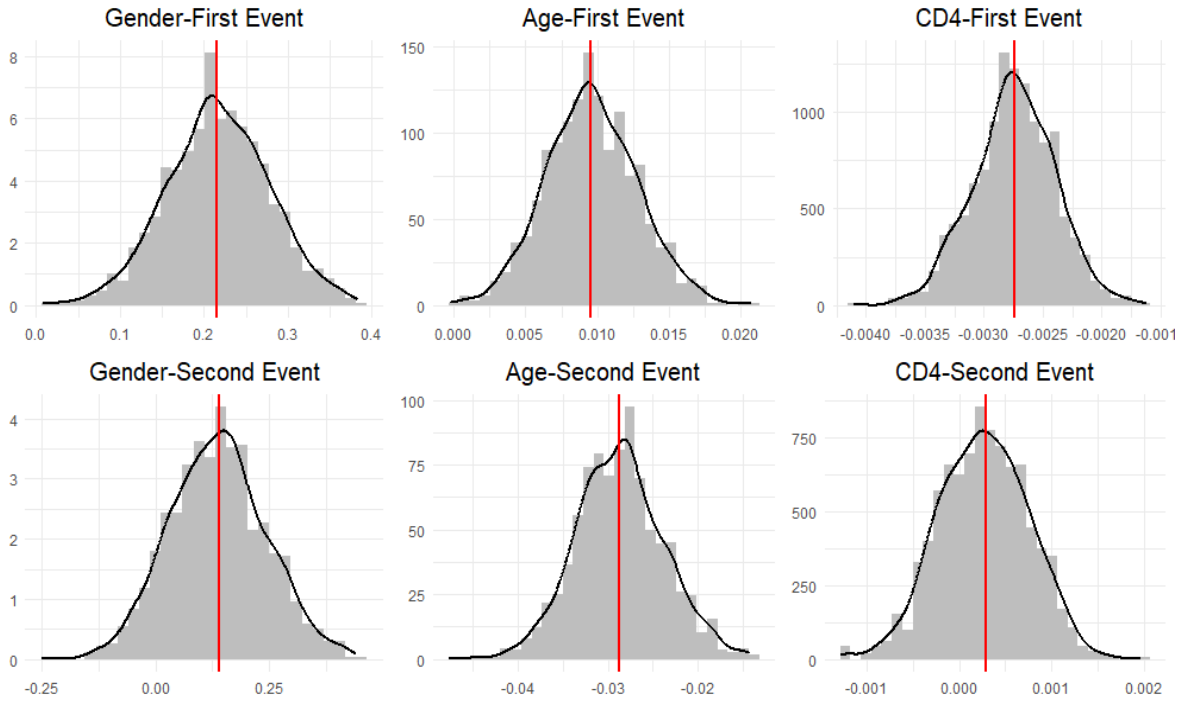
MPPLE third scenario. In addition, the ESE and AMSE are similar in the MPPLE method and in the Bayesian are almost the same. Both the ESE and AMSE are bigger in the MPPLE method. In Table 10, the target value belongs to all 95% confidence intervals expected from the Bayesian third scenario. The ESE and AMSE are the same in all three scenarios. In Table 11, the Bias, the MSE, and the relative absolute error are lower in the Bayesian Method. The target value belongs to the 95% confidence interval. The ESE and AMSE are similar in the MPPLE method but in the Bayesian method are different. Also, the ESE and AMSE are bigger in the MPPLE method. In Table 12, the Bias, the MSE, and the relative absolute error are bigger in the Bayesian method. The Age coefficient in each event does not belong to the 95% confidence interval of the coefficients mean in the third scenario. There is a difference in the ESE and the AMSE in the Bayesian method and in the MPPLE the ESE and AMSE are bigger than the Bayesian one. In Table 13, the Bias, MSE, and relative absolute error are bigger in the MPPLE method but in both methods, they are bigger than 2%. The ESE and AMSE are similar in both methods but the ESE and AMSE are bigger in the MPPLE method. The Bayesian third scenario does not efficiently estimate the CD4 coefficient in relation to the MPPLE. In the Bayesian method, the difference between the ESE and AMSE is bigger at the gender coefficient and almost zero at the CD4 coefficient. In Tables 14,15 and 16 the Bayesian method does not efficiently estimate the scale and shape parameters because in most scenarios they do not belong in the 95% confidence interval of scale and shape parameters. All in all, the ESE and AMSE in the MPPLE method are bigger than the Bayesian one. Nevertheless, the ESE and AMSE are equal in the MPPLE method but in the Bayesian, they are not. It seems that both method produces valid results, but the Bias in the Bayesian method is less than the MPPLE when the assumptions of the Bayesian method are valid.

Graph 13: Histograms of MPPLE method simulations for the first scenario.



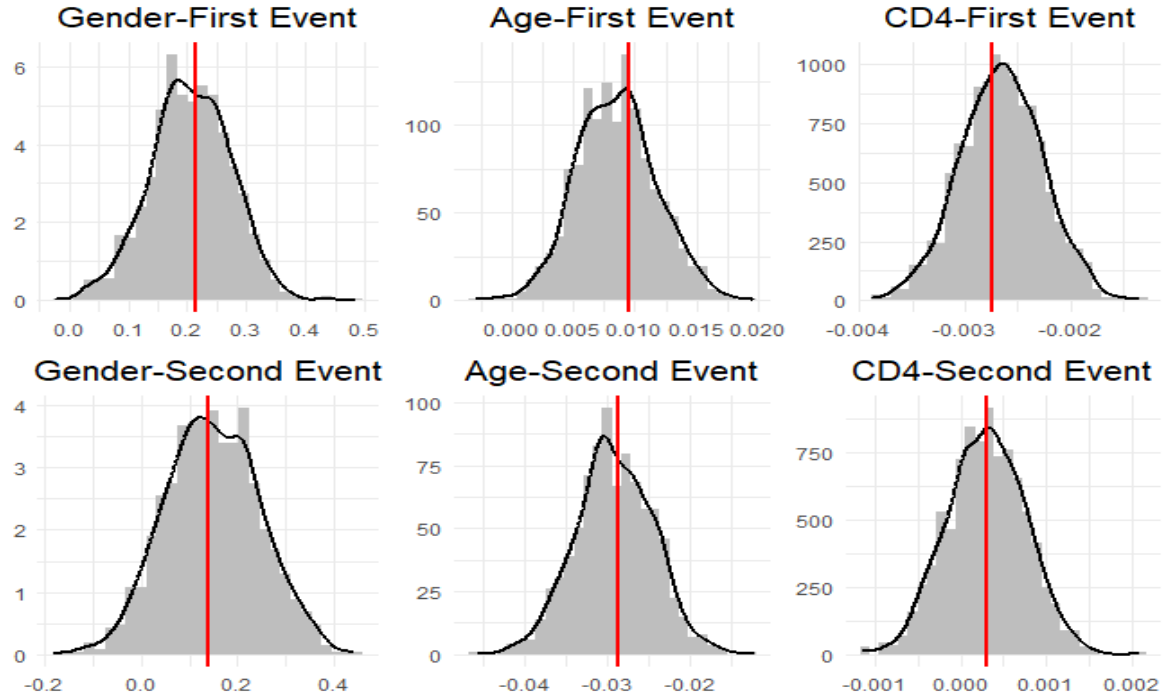
*The red line is the true value of the coefficient.

Graph 14: Histograms of MPPLE method simulations for the second scenario.



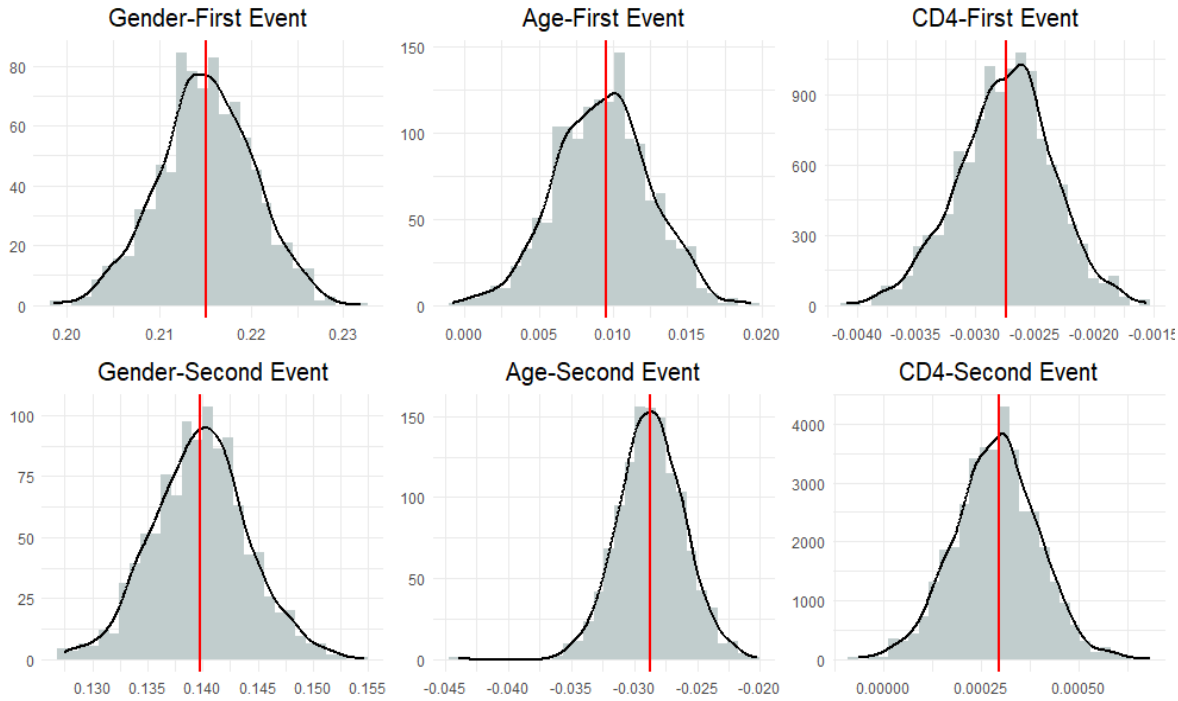
*The red line is the true value of the coefficient.

Graph 15: Histograms of MPPLE method simulations for the third scenario.



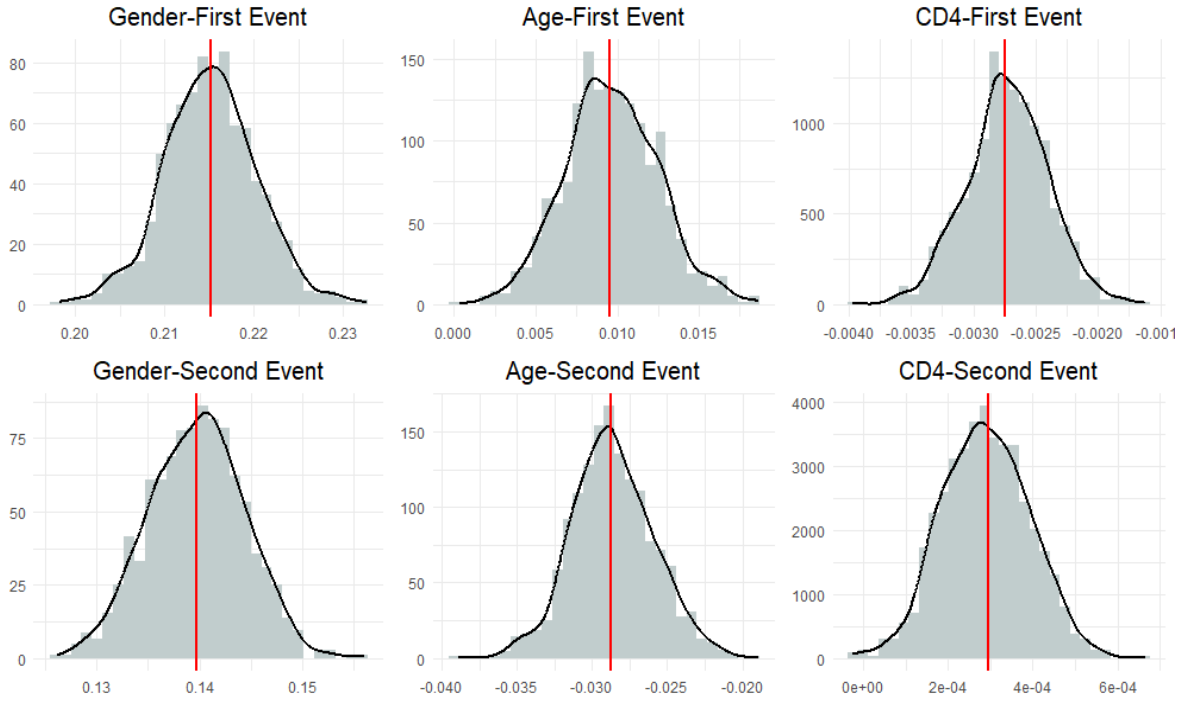
*The red line is the true value of the coefficient.

Graph 16: Histograms of Bayesian method simulations for the first scenario.



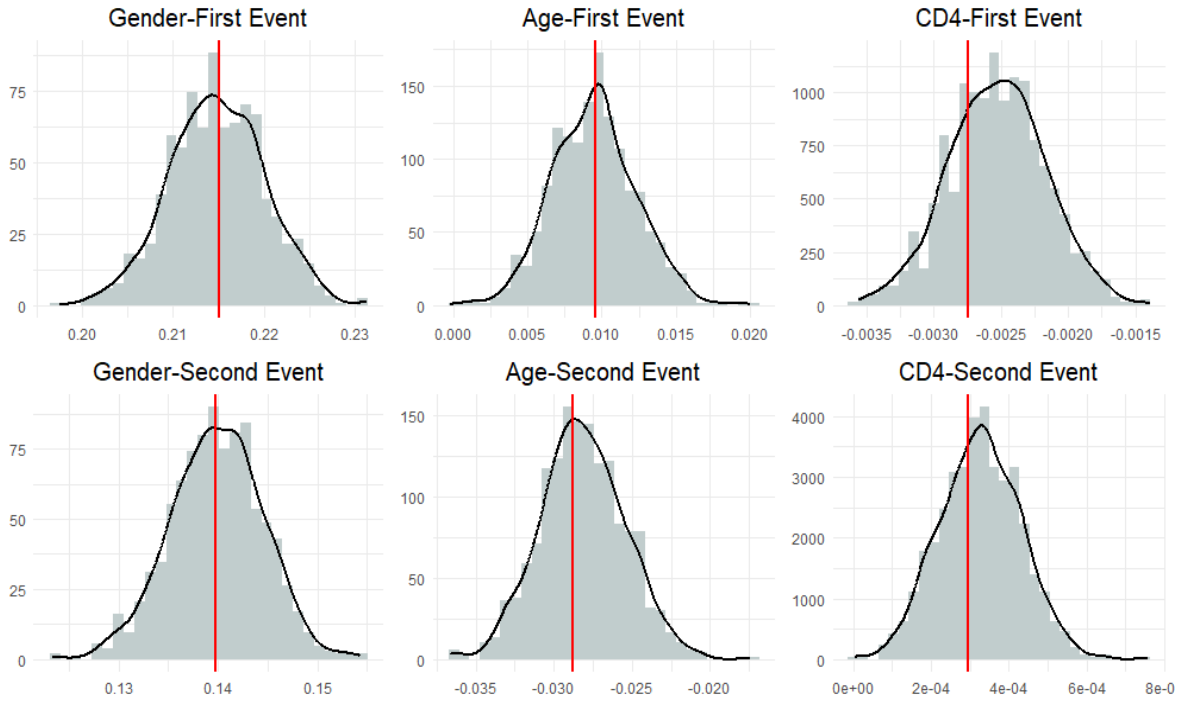
*The red line is the true value of the coefficient.

Graph 17: Histograms of Bayesian method simulations for the second scenario.



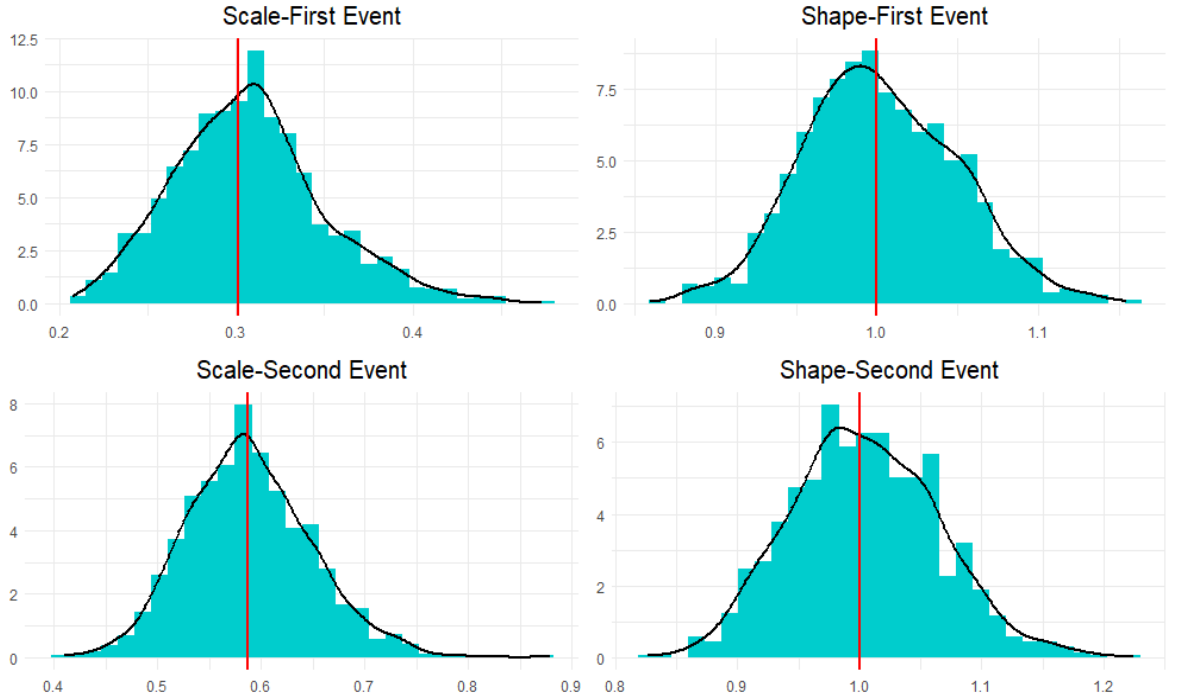
*The red line is the true value of the coefficient.

Graph 18: Histograms of Bayesian method simulations for the third scenario.



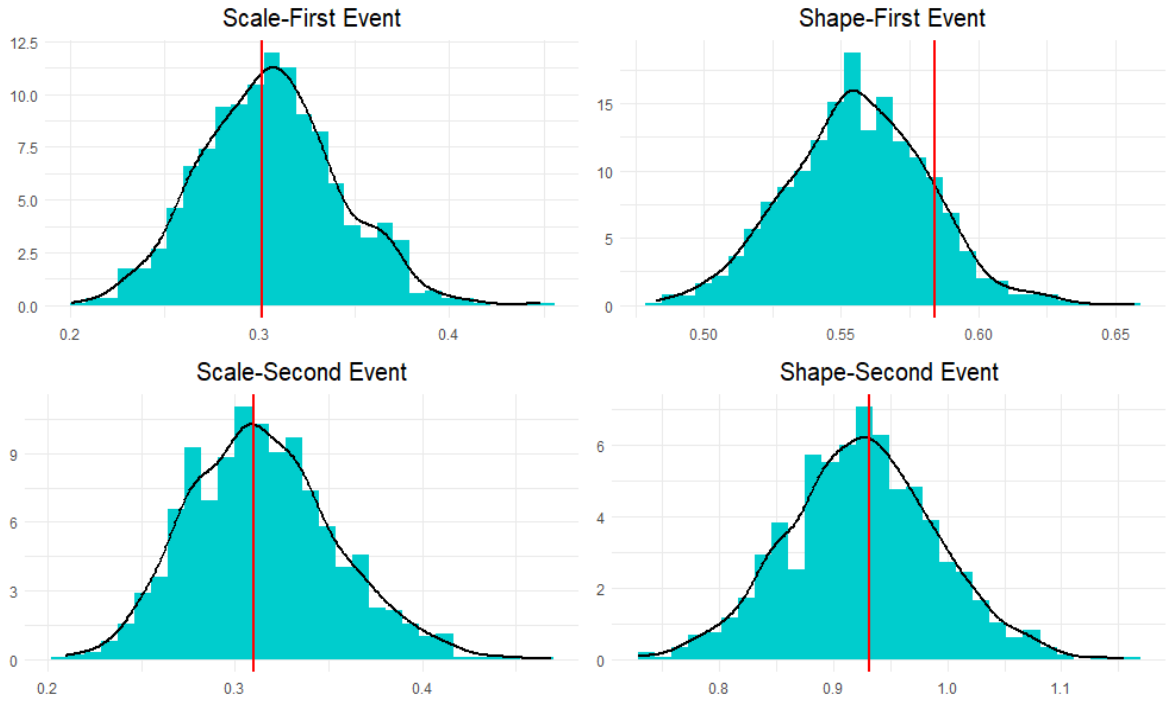
*The red line is the true value of the coefficient.

Graph 19: Histograms of Bayesian method parameter simulations for the first scenario.



*The red line is the true value of the parameter.

Graph 20: Histograms of Bayesian method parameter simulations for the second scenario.



*The red line is the true value of the parameter.

Graph 21: Histograms of Bayesian method parameter simulations for the third scenario.

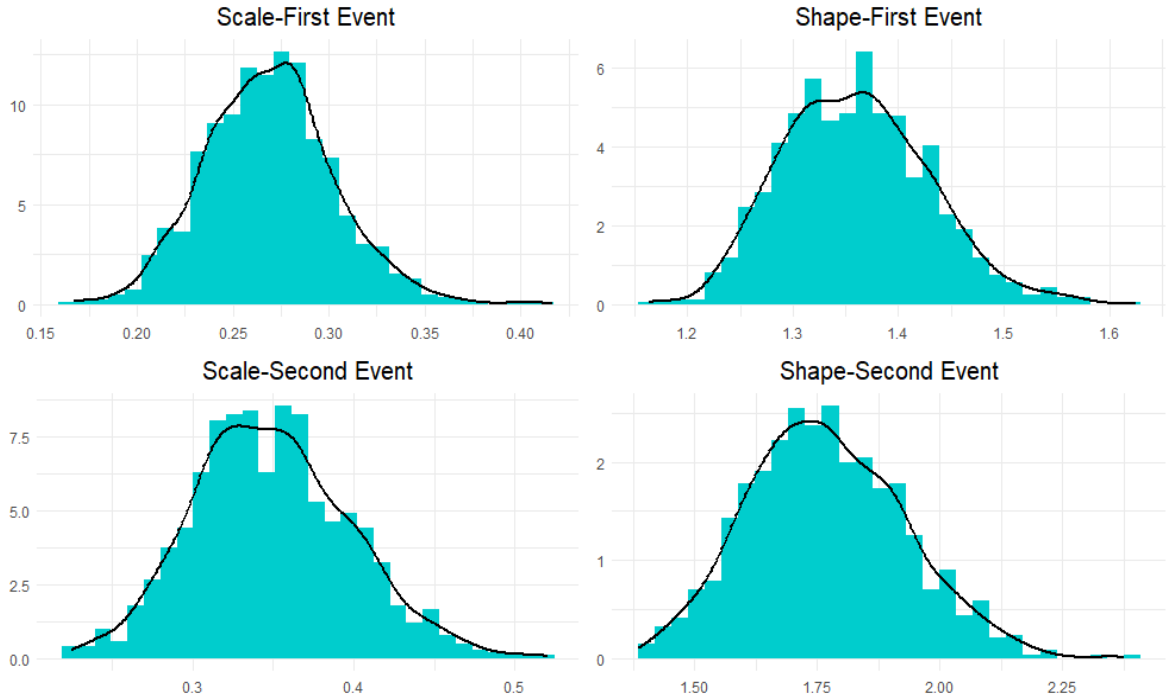


Table 8: Simulations results for the Gender coefficient (0.2151) of the First event.

Scenario	1		2		3	
Method	MPPLE	Bayesian	MPPLE	Bayesian	MPPLE	Bayesian
Bias	0.0007	-0.00002	0.0016	0.00001	-0.0095	-0.0003
MCSE	0.0023	0.0002	0.0019	0.0002	0.0022	0.0002
ESE	0.0721	0.0051	0.0612	0.0052	0.0684	0.0052
MSE	0.0052	0.00003	0.0037	0.00003	0.0048	0.00003
AMSE	0.0726	0.0068	0.0618	0.0066	0.0668	0.0069
RE%	0.328%	0.010%	0.746%	0.004%	4.427%	0.129%
95%QI	(0.066,0.353)	(0.205,0.225)	(0.093,0.336)	(0.204,0.225)	(0.068,0.333)	(0.205,0.225)
95%CI	(0.211,0.220)	(0.215,0.215)	(0.213,0.221)	(0.215,0.215)	(0.201,0.210)	(0.215,0.215)

MPPLE, maximum pseudo-partial likelihood; Bayesian, Bayesian Weibull Competing Risk model with missing events type; MCSE, Monte Carlo Standard estimate of bias; ESE, Empirical Standard Error; MSE, mean square error; AMSE, Average Model Standard error; RE %, Relative absolute error %; 95QI, the 0.025 and 0.975 quantile of the simulated coefficients-parameters; 95%CI the 95% Confidence Interval of the coefficients-parameters mean.

Table 9: Simulations results for the Age coefficient (0. 0095) of the First event.

Scenario	1		2		3	
Method	MPPLE	Bayesian	MPPLE	Bayesian	MPPLE	Bayesian
Bias	-0.0001	-0.0003	0.0001	0.0001	-0.0010	-0.0001
MCSE	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
ESE	0.0036	0.0032	0.0032	0.0029	0.0032	0.0028
MSE	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
AMSE	0.0036	0.0024	0.0031	0.0022	0.0033	0.0024
RE%	1.073%	2.706%	0.809%	0.937%	10.695%	0.772%
95%QI	(0.002,0.017)	(0.003,0.015)	(0.004,0.016)	(0.004,0.015)	(0.002,0.015)	(0.004,0.015)
95%CI	(0.0092,0.0097)	(0.0091,0.0095)	(0.0094,0.0098)	(0.0094,0.0098)	(0.0083,0.0087)	(0.0093,0.0096)

MPPLE, maximum pseudo-partial likelihood; Bayesian, Bayesian Weibull Competing Risk model with missing events type; MCSE, Monte Carlo Standard estimate of bias; ESE, Empirical Standard Error; MSE, mean square error; AMSE, Average Model Standard error; RE %, Relative absolute error %; 95QI, the 0.025 and 0.975 quantile of the simulated coefficients-parameters; 95%CI the 95% Confidence Interval of the coefficients-parameters mean.

Table 10: Simulations results for the CD4 coefficient (-0.0027) of the First event.

Scenario	1		2		3	
Method	MPPLE	Bayesian	MPPLE	Bayesian	MPPLE	Bayesian
Bias	-0.00004	-0.000006	<0.000001	0.00001	0.0001	0.0002
MCSE	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
ESE	0.0004	0.0004	0.0003	0.0003	0.0004	0.0004
MSE	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001
AMSE	0.0004	0.0003	0.0004	0.0003	0.0004	0.0003
RE%	0.144%	0.211%	0.005%	0.574%	3.288%	7.881%
95%QI	(-0.004,-0.002)	(-0.004,-0.002)	(-0.003,-0.002)	(-0.003,-0.002)	(-0.003,-0.002)	(-0.003,-0.002)
95%CI	(-0.0028,-0.0027)	(-0.0028,-0.0027)	(-0.0028,-0.0027)	(-0.00274,-0.0027)	(-0.0027,-0.0026)	(-0.0026,-0.0025)

MPPLE, maximum pseudo-partial likelihood; Bayesian, Bayesian Weibull Competing Risk model with missing events type; MCSE, Monte Carlo Standard estimate of bias; ESE, Empirical Standard Error; MSE, mean square error; AMSE, Average Model Standard error; RE %, Relative absolute error %; 95QI, the 0.025 and 0.975 quantile of the simulated coefficients-parameters; 95%CI the 95% Confidence Interval of the coefficients-parameters mean.

Table 11: Simulations results for the Gender coefficient (0. 1398) of the Second event.

Scenario	1		2		3	
Method	MPPLE	Bayesian	MPPLE	Bayesian	MPPLE	Bayesian
Bias	-0.0005	0.000003	-0.0035	-0.00003	0.0089	0.0001
MCSE	0.0028	0.0001	0.0033	0.0001	0.0031	0.0001
ESE	0.0880	0.0047	0.1055	0.0047	0.0976	0.0047
MSE	0.0077	0.00002	0.0111	0.00001	0.0096	0.00002
AMSE	0.0898	0.0055	0.1034	0.0058	0.0974	0.0059
RE%	0.348%	0.002%	2.500%	0.019%	6.384%	0.070%
95%QI	(-0.037,0.308)	(0.131,0.148)	(-0.063,0.347)	(0.131,0.148)	(-0.036,0.340)	(0.130,0.149)
95%CI	(0.134,0.145)	(0.139,0.140)	(0.130,0.143)	(0.139,0.140)	(0.143,0.155)	(0.1396,0.1401)

MPPLE, maximum pseudo-partial likelihood; Bayesian, Bayesian Weibull Competing Risk model with missing events type; MCSE, Monte Carlo Standard estimate of bias; ESE, Empirical Standard Error; MSE, mean square error; AMSE, Average Model Standard error; RE %, Relative absolute error %; 95QI, the 0.025 and 0.975 quantile of the simulated coefficients-parameters; 95%CI the 95% Confidence Interval of the coefficients-parameters mean.

Table 12: Simulations results for the Age coefficient (-0.02877) of the Second event.

Scenario	1		2		3	
Method	MPPLE	Bayesian	MPPLE	Bayesian	MPPLE	Bayesian
Bias	-0.00003	0.0001	0.0001	0.0001	-0.0004	0.0006
MCSE	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001
ESE	0.0043	0.0026	0.0049	0.0027	0.0047	0.0027
MSE	0.00002	0.00001	0.00001	0.00001	0.00002	0.00001
AMSE	0.0042	0.0022	0.0048	0.0024	0.0046	0.0024
RE%	0.312%	0.401%	0.301%	0.493%	1.489%	2.111%
95%QI	(-0.037,-0.020)	(-0.034,-0.023)	(-0.038,-0.019)	(-0.034,-0.023)	(-0.038,-0.020)	(-0.033,-0.023)
95%CI	(-0.0291,-0.0285)	(-0.029,-0.028)	(-0.0290,-0.0284)	(-0.0288,-0.0285)	(-0.0295,-0.0289)	(-0.0283,-0.0280)

MPPLE, maximum pseudo-partial likelihood; Bayesian, Bayesian Weibull Competing Risk model with missing events type; MCSE, Monte Carlo Standard estimate of bias; ESE, Empirical Standard Error; MSE, mean square error; AMSE, Average Model Standard error; RE %, Relative absolute error %; 95QI, the 0.025 and 0.975 quantile of the simulated coefficients-parameters; 95%CI the 95% Confidence Interval of the coefficients-parameters mean.

Table 13: Simulations results for the CD4 coefficient (0. 000296) of the Second event.

Scenario	1		2		3	
Method	MPPLE	Bayesian	MPPLE	Bayesian	MPPLE	Bayesian
Bias	-0.00002	-0.00001	-0.00003	-0.00001	0.00001	0.00003
MCSE	0.00001	0.000003	0.00002	0.00001	0.00001	0.000003
ESE	0.0004	0.0001	0.0005	0.0001	0.0005	0.0001
MSE	<0.000001	<0.0000001	<0.000001	<0.000001	<0.000001	<0.000001
AMSE	0.0004	0.0002	0.0005	0.0002	0.0005	0.0002
RE%	5.550%	2.431%	10.070%	2.230%	2.091%	10.445%
95%QI	(-0.0006,0.0011)	(0.00001,0.0005)	(-0.0007,0.0012)	(0.0001,0.0005)	(-0.0006,0.0011)	(0.0001,0.0005)
95%CI	(0.00025,0.00031)	(0.00028,0.00030)	(0.00024,0.00030)	(0.00028,0.0003)	(0.00027,0.00033)	(0.00032,0.00033)

MPPLE, maximum pseudo-partial likelihood; Bayesian, Bayesian Weibull Competing Risk model with missing events type; MCSE, Monte Carlo Standard estimate of bias; ESE, Empirical Standard Error; MSE, mean square error; AMSE, Average Model Standard error; RE %, Relative absolute error %; 95QI, the 0.025 and 0.975 quantile of the simulated coefficients-parameters; 95%CI the 95% Confidence Interval of the coefficients-parameters mean.

Table 14: The Evaluation of Scale and Shape parameters of the Bayesian Weibull Model in the first scenario.

Event	Death		Disengagement	
Parameter	Scale	Shape	Scale	Shape
True Value	0.3012	1	0.5873	1
Bias	-0.0059	0.0018	0.0014	0.0040
MCSE	0.0013	0.0015	0.0019	0.0019
ESE	0.0426	0.0474	0.0604	0.0603
MSE	0.0019	0.0022	0.0036	0.0036
AMSE	0.0327	0.0468	0.0551	0.0586
RE%	1.9680%	0.1836%	0.231%	0.396%
95%QI	(0.233,0.399)	(0.911,1.097)	(0.483,0.719)	(0.893,1.126)
95%CI	(0.304,0.310)	(0.999,1.005)	(0.585,0.592)	(1.000,1.008)

MCSE, Monte Carlo Standard estimate of bias; ESE, Empirical Standard Error; MSE, mean square error; AMSE, Average Model Standard error; RE %, Relative absolute error %; 95QI, the 0.025 and 0.975 quantile of the simulated coefficients-parameters; 95%CI the 95% Confidence Interval of the coefficients-parameters mean.

Table 15: The Evaluation of Scale and Shape parameters of the Bayesian Weibull Model in the second scenario.

Event	Death		Disengagement	
Parameter	Scale	Shape	Scale	Shape
True Value	0.3012	0.5840	0.3096	0.9310
Bias	-0.0044	-0.0280	0.0050	-0.0059
MCSE	0.0012	0.0008	0.0012	0.0021
ESE	0.0365	0.0258	0.0392	0.0654
MSE	0.0013	0.0014	0.0016	0.0043
AMSE	0.0296	0.0198	0.0342	0.0610
RE%	1.454%	4.7912%	1.599%	0.6297%
95%QI	(0.236,0.376)	(0.505,0.608)	(0.246,0.398)	(0.799,1.060)
95%CI	(0.303,0.309)	(0.554,0.558)	(0.312,0.317)	(0.921,0.929)

MCSE, Monte Carlo Standard estimate of bias; ESE, Empirical Standard Error; MSE, mean square error; AMSE, Average Model Standard error; RE %, Relative absolute error %; 95QI, the 0.025 and 0.975 quantile of the simulated coefficients-parameters; 95%CI the 95% Confidence Interval of the coefficients-parameters mean.

Table 16: The Evaluation of Scale and Shape parameters of the Bayesian Weibull Model in the third scenario which data were simulated from a Gompertz distribution.

Event	Death		Disengagement	
Parameter	Scale	Shape	Scale	Shape
True Value	0.1759	0.5840	0.1974	0.9310
Estimation	0.269	1.359	0.349	1.768
ESE	0.0338	0.0693	0.0489	0.1593
AMSE	0.0280	0.0616	0.0414	0.1289
95%QI	(0.207,0.338)	(1.236,1.507)	(0.262,0.450)	(1.474,2.096)
95%CI	(0.267,0.271)	(1.355,1.364)	(0.346,0.352)	(1.758,1.778)

True Value, the actual value at which the data were created; ESE, Empirical Standard Error; AMSE, Average Model Standard error %;95QI, the 0.025 and 0.975 quantile of the simulated parameters; 95%CI the 95% Confidence Interval of the parameters mean.

5.2.3 Predictive model

In general, the true from of the predictive probability model is [31]:

$$log\frac{p_1}{1-p_1} = log\frac{h_1(t)}{h_2(t)}$$

For every scenario, the hazard functions are different. For the first scenario which the hazard function is fixed as time changes, the true from of the model is

$$log\frac{p_1}{1-p_1} = log\frac{h_1(t)}{h_2(t)} =$$

$$= log\frac{\exp(b_{11}\times Gender + b_{12}\times Age + b_{13}\times CD4)\times scale1}{\exp(b_{21}\times Gender + b_{22}\times Age + b_{23}\times CD4)\times scale2} =$$

$$= log\frac{scale1}{scale2} + (b_{11}-b_{21})\times Gender + (b_{12}-b_{22})\times Age + (b_{13}-b_{23})\times CD4$$

For the second scenario, the general linear model has the form :

$$log\frac{p_1}{1-p_1} = log\frac{h_1(t)}{h_2(t)} =$$

$$= log\frac{\exp(b_{11}\times Gender + b_{12}\times Age + b_{13}\times CD4)\times scale1\times shape1\times t^{shape1-1}}{\exp(b_{21}\times Gender + b_{22}\times Age + b_{23}\times CD4)\times scale2\times shape2\times t^{shape2-1}}$$

$$= log\frac{scale1\times shape1}{scale2\times shape2} + (b_{11}-b_{21})\times Gender + (b_{12}-b_{22})\times Age + (b_{13}-b_{23})\times CD4 + (shape1-shape2)\times logt$$

For the third scenario. The true form of the general linear model is:

$$log\frac{p_1}{1-p_1} = log\frac{h_1(t)}{h_2(t)} =$$

$$= log\frac{\exp(b_{11}\times Gender + b_{12}\times Age + b_{13}\times CD4)\times scale1\times \exp(shape1\times t)}{\exp(b_{21}\times Gender + b_{22}\times Age + b_{23}\times CD4)\times scale2\times \exp(shape2\times t)}$$

$$= log\frac{scale1}{scale2} + (b_{11}-b_{21})\times Gender + (b_{12}-b_{22})\times Age + (b_{13}-b_{23})\times CD4 + (shape1 - shape2)\times t$$

The hazard function has different form in the third scenario because the time to event follows the Gompertz distribution.

In all three scenario I use the following form for both Bayesian and MPPLE method.

$$log\frac{p_1}{1-p_1} = \theta_0 + \theta_1\times logt + \theta_2\times Gender + \theta_3\times Age + \theta_4\times CD4$$

In the first and second scenarios, this form is the appropriate one; nevertheless, in the third scenario, the form is wrong because I use instead of. In general, in the third scenario, both the Bayesian Weibull and the predictive model are wrong. But in the MPPLE method, only the predictive model is wrong, the Cox regression is totally true because the hazard function under the Gompertz distribution has the proportionality assumption. The next Table represents the actual true form of the logistic models and their estimations both in the MPPLE and the Bayesian method.

The results are satisfying in both methods as are represented in the following Table 17. The estimated predictive models (In both methods) are pretty much near the true one. In the third scenario, it is logical that the true fixed term and the true time coefficient are different from the true fixed term and the natural time coefficient.

Table 17: The comparison between the true and estimated predictive models in both methodologies.

First Scenario					
Covariates	TRUE	MPPLE*	Bias (MPPLE)**	Bayesian	Bias (Bayesian)
1	-0.668	-0.679	-0.011	-0.682	-0.014
$Gender$	0.075	0.078	0.003	0.079	0.004
Age	0.038	0.039	0.001	0.039	0.001
$CD4$	-0.003	-0.003	0	-0.003	0
$logt$	0	-0.001	-0.001	-0.003	-0.003
Second Scenario					
1	-0.494	-0.515	-0.021	-0.512	-0.018
$Gender$	0.075	0.081	0.006	0.082	0.007
Age	0.038	0.039	0.001	0.039	0.001
$CD4$	-0.003	-0.003	0	-0.003	0
$logt$	-0.347	-0.356	-0.009	-0.348	-0.001
Third Scenario***					
1	-0.115	-0.475	-	-0.477	-
$Gender$	0.075	0.057	-0.018	0.059	-0.016
Age	0.038	0.038	0	0.038	0
$CD4$	-0.003	-0.003	0	-0.003	0
$logt$	-	-0.287	-	-0.279	-

*The column MPPLE and Bayesian represent the estimated predictive model (its coefficients are the mean of the 1000 predictive models).

** The columns Bias MPPLE and Bayesian are the estimated coefficients minus the true ones.

*** The true forms and therefore the Bias for them are not given because the actual true form does not have the term but. The true coefficient of t is equal to -0.347. Also, the fixed term is different because of the natural logarithm.

5.3 Sensitivity Analysis

5.3.1 Increasing Prior Variance

The purpose of the first sensitivity analysis is to check how sensitive are the results to the variance of the prior distribution. In other words, it checks if one less informative prior distribution can increase the Bias or MSE and if the scale reduction factor is dependent on it. The first reason is the important one because the deviance of the estimated coefficient informs us how viable and valid the method is especially when the variance of the prior distribution is too low Table 3. The sensitivity analysis is conducted only for the first scenario because of limited time. In order to change the low variance, I multiply both the first and second event's covariance matrixes by 100,000. From, the Bayesian parameters, I change the covariance matrix of the generators, for the first event(death) and the second event (disengagement),the covariance matrix of both of them is equal to $0.00000015\times \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ leading to 0.392 and 0.391 mean acceptance

probability each. For the scale parameter, which is equal to 0.301, the standard deviation of the generator is equal to 0.014 and for the shape parameter, which is equal to one, the standard deviation of the generator is 0.036 resulting in 0.51 and 0.51 mean acceptance probability. The starting points for the scale and shape parameters are 0.301 and one respectively. For the scale parameter of the second event, which is equal to 0.587, the standard deviation of the generator is equal to 0.034 and for the shape parameter, which is equal to one, the standard deviation of the generator is 0.041 resulting in 0.488 and 0.514 mean acceptance probability. The starting points for the scale and shape parameters are 0.587 and one respectively.

Table 17: The evaluation results of the Sensitivity analysis with less informative prior in the first scenario.

Event	Death				
Coefficient	Gender	Age	CD4	Scale	Shape
Bias	-0.0006	-0.0003	-0.00001	0.0071	0.0020
MCSE	0.0002	0.0001	0.00001	0.0015	0.0015
ESE	0.0063	0.0036	0.0004	0.0472	0.0475
MSE	0.00004	0.00001	<0.000001	0.0023	0.0023
AMSE	0.0077	0.0027	0.0004	0.0358	0.0471
RE%	0.265%	2.892%	0.257%	2.353%	0.201%
95%QI	(0.202,0.227)	(0.002,0.016)	(-0.004,-0.002)	(0.227,0.411)	(0.912,1.097)
95%CI	(0.214,0.215)	(0.0090,0.0095)	(-0.00278,-0.00272)	(0.305,0.311)	(0.999,1.005)
Event	Disengagement				
Coefficient	Gender	Age	CD4	Scale	Shape
Bias	-0.0001	0.00003	-0.00004	0.0128	0.0041
MCSE	0.0002	0.0001	0.00001	0.0032	0.0019
ESE	0.0066	0.0042	0.0004	0.1002	0.0608
MSE	0.00004	0.00002	<0.000001	0.0102	0.0037
AMSE	0.0079	0.0030	0.0004	0.0751	0.0591
RE%	0.051%	0.088%	14.596%	2.183%	0.4114%
95%QI	(0.127,0.152)	(-0.036,-0.020)	(-0.0006,0.0010)	(0.424,0.818)	(0.893,1.124)
95%CI	(0.139,0.140)	(-0.029,-0.028)	(0.00023,0.00028)	(0.594,0.606)	(1.000,1.008)

*MCSE, Monte Carlo Standard estimate of bias ;ESE, Empirical Standard Error ;MSE, mean square error ;AMSE, Average Model Standard error ;RE %, Relative absolute error % ;95QI, the 0.025 and 0.975 quantile of the simulated parameters ; 95%CI the 95% Confidence Interval of the mean.

First of all, I want to point out that all real coefficients belong to the mean 95% CI except the scale parameters which are overestimated, and the CD4 in the second event. All true coefficients belong to a 95% quantile interval. The two scale parameters, the Age coefficient in the first event and the second event CD4 coefficient have the biggest relative error in opposition to the Gender coefficient in both events which have the lowest. The Bias, MSE, ESE, and AMSE of the

coefficients are insignificantly bigger than the Bias, MSE, ESE, and AMSE of Table 8 -13 in the first scenario. The results are satisfying, just the Bias, MSE, and RE% have an unimportant rise. Only the CD4 coefficient in the second event has a concerning result with a 14.6% relative error. All in all, the less informative prior does not significantly impact the results except for the CD4 in the second event and the standard error of the coefficients.

6. Discussion

6.1 General

6.1.1 Bayesian Weibull competing Risk with missing cause of failure

The main priority of the master thesis is to describe and model the Bayesian Weibull competing risks with missing event types under MAR assumption. The strategy is to break down this model into two parts. The first part is about modeling the Bayesian Weibull competing risks model which needs a complete dataset. The second one is about imputing the missing dataset utilizing one data augmentation method and one Bayesian asymptotic property. The concept of the specific missing values imputation method is about first imputing the missing dataset and then recalculating the coefficients and their covariance matrix using the ordinary maximum likelihood method. After that simulate the chosen coefficients for this round from a multivariate normal distribution with mean equal to the coefficients and with covariance equal to the coefficient's covariance matrix. Then proceed to the simulation of the desirable Bayesian Weibull competing risks coefficients and parameters after that the same process is re-conducted. The simulation of the 'missing' coefficients is done by utilizing one asymptotic property. The coefficients follow a multivariate normal distribution with mean and a covariance matrix the estimated likelihood coefficients and their covariance matrix, this asymptotic property is valid when the prior is flat, and the number of observations is big enough like this scenario. In practice, the missing values are treated as parameters that are simulated from their posterior predictive probability. The simulated values from the posterior predictive probability are simulated first by simulating the coefficients of the predictive model and then they are 'predicted' from the model. So, the same process is conducted here, first, the simulation of the coefficients is done and then the prediction of the missing values using the simulated coefficients is conducted. The continuous re-prediction of the missing values is conducted because they are unknown parameters, and they are needed for the estimation of the Weibull coefficients and the parameters.

The Bayesian-Weibull-competing-risk part with two events is about estimating two Bayesian Weibull survival models. For each event, the other event is treated as censored; as a result, the same procedure is conducted twice. The Bayesian Weibull survival model is estimated by calculating the coefficients and the scale plus shape parameters. In order to simulate the coefficients, the posterior distribution of the coefficients is essential. The posterior distribution is impossible to calculate because the normalizing constant cannot be estimated; as a result, only the likelihood multiplied by the prior distribution is utilized. For this reason, a Metropolis Hasting method which belongs to Markov Chain Monte Carlo models is chosen instead of a Gibbs one. Particularly, a Metropolis Hasting with Random Walk is chosen because it is easier to program, and all parameters are easily tuned. For instance, the variance of the generator(normal distribution) is tuned by the acceptance probability and the mean of the generator is the previous value. Also, the symmetric property of the generator eliminates its density in the probability acceptance. As a result, only, the likelihoods and the prior of the parameters are

left in the formula of the acceptance probability. Because of that the Metropolis Hasting algorithm with Random Walk is the convenient choice. Nevertheless, the Metropolis-Hastings with Random Walk produces simulations that have high autocorrelation. In general, this Bayesian Weibull competing risk method with missing events is a method that first deals with the missing cause of failure and then uses a Bayesian methodology to estimate the Weibull coefficients
plus, the scale and shape coefficients.

The advantages and the disadvantages of the Bayesian Weibull Competing Risk with missing cause of failure are numerous. To begin with, the first both advantage and disadvantage is that the method utilizes Bayesian Statistics. Bayesian Statistics are better when the sample size is small and when there is specific information for the prior distribution., another advantage is that Bayesian statistics uses the probability theory directly. They use the Bayesian theorem in all aspects of models and formulas. In every problem, there is a prior distribution, a likelihood, and a posterior distribution. All the information is in the posterior distribution or the simulated values of the posterior. Nevertheless, Bayesian Statistics have the problem of efficiently defining one good prior or one prior which makes sense. In this master thesis, all the information about the prior distribution is taken from the [16]. As a consequence, the prior choice that has been made is a sensible option based on a particular paper. However, this prior knowledge sometimes has a huge impact on the coefficients. Particularly, the prior distribution that is used is very informative and can manipulate the results but because the number of observations of each dataset is more than 6000 observations, the impact of the prior distribution becomes meaningless, and the results are more like the corresponding frequentist approach. One huge disadvantage of Bayesian statistics is that it requires high statistical knowledge which this knowledge sometimes does not be taught in the bachelor's or master's degree. Apart from that it also requires a great knowledge of programming and a good computer to run hundreds of thousands of simulations. In this master thesis, the number of simulations in each scenario, in each data, and in each chain is too low. Consequently, the mean scale reduction factor is above one in several situations (see Appendix) and that indicates that convergence might not have been achieved. If the real results had been unknown, a such big scale reduction factor would have been alarming. In practice, the number of simulations in each chain should be higher and the initial value of each chain should be unequal. Another, disadvantage of Bayesian statistics is that the convergence of one or more coefficients needs too many simulations. In addition, in order to conduct one good Bayesian study various sensitivity analyses which are related to the prior distribution are necessary to be conducted. The Weibull competing risk is a way to model data using the Weibull formula or just apply a Weibull model in the baseline hazards. This Weibull model has two requirements first the proportionality of hazards and then the baseline hazard must have one specific form. Because of that, the coefficient estimations in the third scenario are bad because the data for the third have been simulated using the Gompertz distribution. Nevertheless, there are various ways to model the baseline hazard, for instance, one very popular way is the accelerated failure time model. The Weibull parametric model can also be described as one accelerated failure time model. Other AFT models are the log-logistic, the lognormal, the gamma, the

inverse Gaussian, and more. Those AFT models don't require the proportionality property and their hazards are not monotonous like the Weibull one which is a straight line. In each of those models, the baseline hazard has a different form and for each hazard form, the survival functions are derived. Using the survival and hazard function, one can derive the likelihood and by assigning priors to the coefficients, can create for example one Bayesian log-logistic competing risk model with missing cause of failure. So, an analyst can try different models and choose which of them best fits the data. One should consider that the AFT models utilize distinct forms of parameterizing the coefficients. This distinct form models the log time and the covariates in a linear way. In general, after having one hazard and its survival function, one can easily incorporate them in the likelihood and just use the same Metropolis-Hastings with Random Walk. One another advantage of the Bayesian Weibull Competing Risk with the missing cause of failure is that the missing cause of failures is treated like parameters that need to be estimated. Consequently, the imputations of the missing cause of failure are practically a Bayesian problem that is solved and applied inside the Metropolis-Hastings loop. The Bayesian Weibull competing risks model with missing events type uses a complete dataset to estimate the coefficients and the algorithm is the same as the corresponding Weibull competing risks without the missing events. One another disadvantage is that the probability model referred to missing values must be efficiently and meticulously examined for its predictable capability. In other words, it is essential that the predictive model must be valid, and its predictions are on average true. If this predictive probability model had not been adequately modeled, all the predictions would have been wrong or misleading. If the model is too simple, it would lose information and if the model is too saturated, the predictions would be wrong because of the overfitting. Also, the predictive model is restricted to a general linear model because all the papers indicate that. This master thesis is a simulation study, and the form of the predictive probability is already known; therefore, the form of the estimated predictive models is the same for the first and second scenarios but in every estimated predictive model there is uncertainty about the coefficients being approximately near the true ones. Finally, the variable selection which this master thesis has not been involved with is a hard and tedious procedure. Because the coefficients for specific combinations of parameters have to be estimated in order to compare those to another estimated model. This procedure can absorb a significant amount of time because in order to have valid results for each coefficient a convergence has to be achieved and that can take time because it might need more simulations or time to tweak the initial parameters.

6.1.2 Maximum Pseudo-Partial-Likelihood Estimation method

This method has not been described in a theoretical base. In this master thesis, only the methodology to apply this MPPLE method has been described. The actual reason why this method works is in the [16] and has not been described here because this master thesis purpose is to describe and compare the Bayesian Weibull competing risk with missing event types. Practically, the methodology and the application of the MPPLE method are far simpler than the corresponding Bayesian model. In terms of accuracy and validity, it is the same as the Bayesian

Weibull competing risk with missing cause of failure. The MPPLE method is a more straightforward method because it does not require strong programming and unique statistical knowledge (like Bayesian statistics). Also, it does not require a significant time to just calculate one coefficient estimation and its standard deviation like the Bayesian method. In addition, this method utilizes packages that already exist in R like the survival package because this method is basically a weighted Cox regression. The problem with this method is that the standard error of each coefficient needs a bootstrap methodology. The initial sample is resampled with replacement and for each new-replaced data, a coefficient is calculated. This procedure is iterated for one hundred or more times and the standard deviation of the bootstrapped coefficients is the standard error of it. This method also suffers the same problem with the prediction model as the Bayesian method. Practically, this method requires only the probabilities of the event being the first one if there are two events. The Bayesian method simulates the event with respect to the probability of the event being the first event (or the event of interest). The missing observations are duplicated, and in the first bunch their weights are equal to the probability of the missing event being the first one and in the second bunch, their weights are equal to the probability of the missing event being the second (or 1 minus the previous probability). Then the initial complete dataset with weights being equal to 1 and the rest imputed dataset, whose length is two times the length of the initial missing dataset, have weights equal to the probabilities. Therefore, if the prediction model is wrong or misleading the weights can be misleading too. Also, the variable selection is more conveniently conducted than the Bayesian method because ordinary variable selection methods are required due to the weighted Cox implementation. The problem with this methodology is, that it is appropriately and scientifically applied to the Cox model [16]. Therefore, if the proportionality of the hazards is not valid then the results are not precisely true, but they are true on average. It's like the typical problem of proportionality. So, if the proportionality is not applied then various mechanics like the time covariate or the stratified Cox regression. Nevertheless, if those strategies cannot remedy the situation, there is nothing more to do. In the Bayesian Weibull competing risk model with missing events type, if the Weibull model does not appropriately fit the data, one can change it to the log-logistic, inverse gamma, and more … The MPPLE method is faster, and its results are more accurate and valid because there is not an aspect of convergence. In all tables, the results are adequate enough which indicates that this method is a solid one. Another disadvantage of this method is that it requires more samples than the Bayesian one.

6.2 Concluding Remarks

6.2.1 Advantages of the methods and study

The main advantage of the study is that compares two methods that try to deal with the missing cause of failure in competing risk. Those methods are directly compared, the Bayesian method has less Bias, MSE, and relative absolute error in the first and second scenario but it loses in the third one which the data are simulated via the Gompertz distribution. The AMSE and the ESE of the MPPLE method are similar, but this is not applied to the Bayesian method. Nevertheless, the AMSE and ESE are bigger in the MPPLE method in relation to the Bayesian. In general, both methods face the competing risk problem with missing cause of failure adequately and they handle the missing cause problem efficiently. Also, this study covers a space in the bibliography because there is no study that compares a frequentist and a Bayesian approach in competing risk with the missing cause of failure. Also, in the bibliography, a model like the Bayesian Weibull competing risk with missing cause of failure has not been described and programmed. This master thesis shows thoroughly how the Bayesian Weibull Competing risk is derived the problems that may show up and how to deal with them. Another advantage of the study is that the Bayesian method is tested in 3 scenarios and the results indicate that if the hazard function is not a straight line or just the time does not follow a Weibull distribution then the results might be wrong or misleading. Also, the results of the MPPLE method are very satisfying because this method has trustworthy results in all 3 scenarios. In addition, one sensitivity analysis has been conducted. The sensitivity analysis indicates that if the prior becomes less informative, the method does not lose its robustness, and the variability of the prior is not correlated with the validity of the results. Finally, one can compare how those methods handle the missing values. Particularly, the two estimated methods are compared with the true probability predictive model, and with scenario 3 one can immediately understand the impact of a good predictive model. In the first and second scenarios, the predictive model is efficiently estimated in both methods. In the third scenario, the form of an estimated predictive model is wrong but the coefficients of the Gender, Age, and CD4 are efficiently estimated. All in all, both methods adequately handle the problem of competing causes of failure with missing causes of failure especially when the assumptions are applied. One should take care that the MPPLE has fewer assumptions than the Bayesian method.

6.2.2 Disadvantages of the methods and the study

The disadvantages of the study are related to the Bayesian Weibull Competing risk with missing cause of failure. Firstly, the number of simulations in chains is too low because of the low computation power and that is a problem. In order to handle this problem, I assign the initial values as the target coefficients; as a result, all chains in one scenario -event start from the same values which are the desirable target values. The low number of simulations and the initial values being the target one cause a big mean scale reduction factor. I have assigned the initial values as equal to the target coefficients because I wanted to somehow accelerate the convergence. The big mean scale reduction factor indicates that convergence has

not been achieved in some variables. Also, the scale reduction factor is a good metric of convergence only if the initial values are different [18]. The combination of an old computer and the big number of datasets; namely, there are 1000 datasets in every 3 scenarios which is practically 3000 datasets. For a dataset often more than 1000 simulations are needed for each chain and apart from that more than 4 chains are essential with different starting points but in this master thesis, only two chains with the same initial point are used. The number of simulations, chains, and starting points were impossible to change because if I change the number of simulations and chains I will need far more time than I need now and If I change the initial values, I will need far more simulations and chains. Furthermore, the burn-in period is relatively small (500 simulations) and I take 1 simulation per 25 because I want to lower the correlation between the chosen simulations. Another disadvantage is that analysts suggest writing Bayesian programs in other statistical software like WinBUGS or Stan and more. In addition, the CD4 should be simulated from a Poisson distribution, because the CD4 are cells per μ1, and not from a normal distribution but this is practically not a big concern because it does not impact the results.

6.2.3 Some final thoughts

To begin with, both the MPPLE method and the Bayesian Weibull Competing risk with missing cause of failure give valid and accurate results but I am more satisfied with the results of the MPPLE method, and I totally suggest it because it is more straightforward, convenient, and faster and it does not require a strong computer and tuning endeavor which sometimes in Bayesian statistics is a tedious and monotonous procedure. The Bayesian method is more sensitive to the assumption than the MPPLE method. Nevertheless, the Bayesian method produces better results when the assumptions are applied. Only two issues might arise first the predictive model should be meticulously described and the second one is the adequacy of the Cox regression method. In some situations, the proportionality does not stand and therefore a stratified Cox or time-coefficient term is necessary. In addition, The Bayesian Competing risk with missing cause of failure is more flexible because other models than the Weibull one can be used such as log-logistic in comparison to the MPPLE methodology which only a Cox regression can be applied. Also, the missing cause of failure might be handled more appropriately in the Bayesian method because the missing cause of failure is treated like a parameter which in the end is theoretically converged to the true value. Both methods can use the same predictive algorithm, but the Bayesian method can enhance its performance because of the Bayesian methodology. Apart from the previous, in this simulation study, the scale and shape of the first and the second scenarios are almost always overestimated and underestimated respectively. The true values of the scale and shape parameters are outside the 95% confidence intervals of the mean in most cases. In the third scenario, the scale and shape parameters of the Bayesian Weibull competing risk model with missing cause of failure are outside the 95% confidence interval of the mean and that is a sensible result because the Bayesian Weibull method is the wrong method and theoretically does not efficiently fit to the data. From the second sensitivity analysis which I simulate more values inside the chains, the shape parameters in both events are underestimated by a small amount and only in the first event the

scale parameter is inside the 95% confidence interval of the mean. This problem of scale and shape parameters might be caused either by missing or censoring events. Another problem is the scale reduction factor which is bigger than one in the Gender, Age, and Scale coefficient – parameters(Appendix). It seems that it is independent of the scenario and the event and there is a specific pattern between the scaler reduction factor and the variables. In all scenarios and events, the mean scale reduction factor is the biggest in the Age coefficient and the lowest in the CD4 coefficient. Also, the mean scale reduction factor of the scale parameter is always bigger than the shape parameter. Specifically, the problem of a big-scale reduction factor lies in the mixing chains. The mixing chains with high-scale reduction factors, first include other chains with high autocorrelation or with high lag correlation. Second, the chains seem to be stuck in another area, some are in the middle, others up and others down, and third each chain has a different variance. The general picture of bad mixing chains indicates that the problem arises when there is a high correlation and there is a different route of each chain which implies weak or incompetent convergence. The combination of chains with small-scale reduction factors is random and stationary, they are like a cloud. In other words, they are all randomly distributed around a value, and they have the same variance. The results of the second sensitivity analysis indicate that as the number of simulations rises the mean scale reduction factor decreases but only to the Age and scale coefficient – parameter. Therefore, if the number of simulations rises the scale reduction factor might decrease. Apart from that the Gender coefficient has the biggest standard error, the biggest value, and the biggest mean scale reduction factor in opposition to the CD4 covariate which has the lowest value, standard error, and scale reduction factors. From all of those above, it is a good question that arises what will happen If I run more simulations and more chains? How many simulations are necessary in order to acquire appropriate convergence metrics?

I would like to model the Weibull parametric model as an accelerated failure time model and not like I did because If I model it like an AFT, it would be easier to fit another AFT model. With the modeling of AFT models, it would be easier to conduct different simulation scenarios with different baseline hazards that are not monotonous. In other words, one can build a library that can run all types of Bayesian parametric competing risk models like the Weibull, Gompertz, log-logistic, inverse gamma, and more. The purpose of this library would be first the modeling of a Bayesian parametric competing risk model with missing event types, and second which of them best fits the data. The test of the goodness of fit could be the ordinary survival test or the Bayesian test related to the predictive model of time to event (comparing the real and estimated time to the event).

Last but not least, how the Bayesian algorithm behave if there were 3 generators one for each coefficient and not one that is multivariate? In other words, the current generator is a multivariate normal distribution but what if every coefficient had its generator like the scale and shape parameters? It is sure that the procedure would take far more time than it has taken. Theoretically, in the long run, the results of the two types of generators are the same but which of them is better in the short term? Also, the current multivariate generator uses a diagonal covariance matrix which is interpreted that the coefficients are not

correlated. All 3 generators would also produce coefficients that are not correlated. The main difference is that the generation is done simultaneously and not one after the other. With 3 generators every simulated value is tested independently if it fits into the likelihood. Nevertheless, with a multivariate generator all 3 values of the vector are tested simultaneously if they fit the data. What if one has a bigger impact than the other? For example, the CD4 coefficient can manipulate the option if this vector is accepted or not more than the other coefficients.

An analyst can use a different type of predictive model like Random Forest, artificial neural network, lasso or ridge regression (elastic net), k nearest neighbors, Extreme gradient boosting (XGBoost), Adaptive boosting (ADA boost), and more. Also, someone can use different algorithms and, in the end, can combine all of them into one algorithm. This can be done either by a voting procedure for example 2 algorithms suggest death and 1 disengagement; as a result, the final choice is death. Also, they can predict the outcome from various models and then use those results as a covariate in another model. In addition, they can use different types of covariate modelling and then combine all those models two one. The validity and the accuracy of all those methods are done via the k-fold cross-validation or bootstrap validation or leave one out and more. There are many 'intelligent' predictive models that one creative analyst has abundant options. Nevertheless, those methods have not been validated by the 'inferential' community. One issue is that we don't know practically the distribution of the parameters. For example, the random forest has two parameters the number of trees and the number of bootstrapped samples. Also, most of the predictive models do not have a likelihood. So, the problem is that there are no probability models; as a result, they don't have likelihood and their parameters don't have prior. Also, the fact that most of them like XGBoost, artificial neural networks, and more are black boxes. Their interpretation power is relatively small, and they don't answer the question of why one model is better intuitively than another but which model is better. For example, let's consider the typical problem of smoking, the yellow finger and the cancer, the previous predictive models might find the yellow finger is a better predictor than the smoking one because the cross-validation of the model just gives better results.

One another topic is the use of another model but Cox in the MPPLE method. In the paper [3] this method is suggested by applying it to only the Cox regression models. One analyst can research if other models like the Weibull parametric, accelerated failure time models can be used in a relative scenario. For example, one can use a weighted Weibull parametric regression with weights equal to the probabilities of the missing values being the first event. There are so many parametric models that can deal with the non-proportionality of the hazard function. Those accelerated time failure models use hazard functions that are not monotonous such as the log-logistic and the Weibull which is a straight line. They try to better fit the hazard function to the real one. If the real hazard function is a cube that fluctuates around a value, or it changes its monotony dramatically like a second-degree polynomial then the straight line is an inadequate option.

Summary

In survival studies, more than one cause of failure is a frequent phenomenon and sometimes the cause of failure is missing under MAR assumption. Several approaches are suggested some of them are Bayesian or frequentist, parametric or semi-parametric methodologies. The Bayesian Weibull competing risk with missing cause of failure is a model that has not been adequately described in the bibliography. As a consequence, this master thesis first tries to efficiently describe this model and second, compare it with an existing Maximum pseudo-partial estimation method. This Bayesian Weibull competing risk with missing event type model uses the Bayesian methodology to impute the missing cause of failure and estimate the desirable coefficients-parameters. The imputation of the missing cause of failure is conducted by treating those missing observations as parameters. The form of the coefficients - parameters of the model are derived from a Weibull competing risk model and they are estimated via a Bayesian methodology. The Maximum pseudo-partial estimation method is a computationally efficient method in which its coefficients are estimated by a weighted-probability Cox regression model. The simulated data are derived from the statistics of the EA-IeDEA HIV study which in this study a heavy under-reporting issue of the event type is observed. The results of the simulation study indicate that both methodologies effectively handle this issue when the assumptions of the models are applied. Both the proportionality and the Weibull assumption significantly impacts the validity of the results in the Bayesian Weibull model, but in case they are not true, one can use accelerated time failure models instead of a Weibull one to tackle this issue.

Περίληψη

Στις μελέτες ανάλυσης επιβίωσης, αρκετές φορές υπάρχει πάρα πάνω από ένα είδος κινδύνου και μάλιστα αρκετές φορές οι ανταγωνιστικοί κίνδυνοι μπορεί να λείπουν. Αρκετές μέθοδοι έχουν προταθεί όπως για παράδειγμα Μπεϋζιανα και μη, παραμετρικά ή σχεδόν παραμετρικά μοντέλα. Το Μπεϋζιανο Γουεϊμπουλ παραμετρικό μοντέλο με τους ελλιπές ανταγωνιστικούς κινδύνους είναι ένα μοντέλο που δεν έχει περιγράφει επαρκώς στην βιβλιογραφία. Ως αποτέλεσμα, αυτή η διπλωματική εργασία έχει σκοπό να περιγράψει σαφώς και με διευκρινιστικό τρόπο το μοντέλο και να το συγκρίνει με μια μέθοδο που χρησιμοποιεί την μέγιστη μερικώς ψευδό πιθανοφάνεια για να εκτιμήσει τους συντελεστές. Το Μπεϋζιανο Γουεϊμπουλ παραμετρικό μοντέλο με τους ελλιπές ανταγωνιστικούς κινδύνους χρησιμοποιεί την Μπεϋζιανη μεθοδολογία για να γεμίσει τις ελλιπείς τιμές και να εκτιμήσει τις επιθυμητές παραμέτρους. Πιο συγκεκριμένα, για να εισαχθούν οι ελλιπείς τιμές χρησιμοποιώντας την Μπεϋζιανη θεωρεία θεωρούνται ως άγνωστοι παράμετροι που πρέπει να εκτιμηθούν. Οι συντελεστές του μοντέλου και οι διάφοροι παράμετροι κτίζονται από το Γουεϊμπουλ παραμετρικό μοντέλο με ανταγωνιστικούς κινδύνους και τελικά εκτιμώνται με την χρήση της Μπεϋζιανης Στατιστικής. Η μέθοδος που εκτιμά τους συντελεστές με την χρήση της μέγιστης μερικώς ψευδό

πιθανοφάνειας είναι μια γρήγορη και παράλληλα αποτελεσματική μέθοδος που πρακτικά εκτιμάει τους συντελεστές μέσω ενός Κοξ σταθμισμένου γραμμικού μοντέλου. Τα προσομοιωμένα δεδομένα δημιουργούνται από τα στατιστικά μιας ερευνάς που έγινε στην Ανατολική Αφρική η οποία λέγεται EA-IeDEA και έχει σκοπό να μελετήσει τον Ιό της Ανοσοανεπάρκειας του Ανθρώπου, στην συγκεκριμένη έρευνα υπάρχουν πολλοί ασθενείς που τελικά είναι άγνωστο για τους ερευνητές αν πεθάναν ή απλά έφυγαν από την έρευνα (δηλαδή ζούνε). Τα αποτελέσματα από την έρευνα προσομοίωσης έδειξαν ότι και οι δυο μέθοδοι είναι αρκετά αποτελεσματική και μπορούν με ευκολία να αντιμετωπίσουν παρόμοια προβλήματα όταν ειδικά ισχύουν οι υποθέσεις του κάθε μοντέλου. Η αξιοπιστία και η εγκυρότητα του μοντέλου εξαρτάται πολύ από την υπόθεση της αναλογικότητας και από την επιθυμητή Γουεϊμπουλ μορφή των κινδύνων , σε περίπτωση όμως που τα αποτελέσματα είναι αμφιλεγόμενα και καθόλου αποτελεσματικά τότε μια καλή λύση είναι κάποιος να χρησιμοποιήσει ένα άλλο παραμετρικό μοντέλο αντί του Γουεϊμπουλ όπως για παράδειγμα ένα μοντέλο επιταχυνόμενου χρόνου.

Reference

1. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation. 2016 Feb 9;133(6):601-9. doi: 10.1161/CIRCULATIONAHA.115.017719. PMID: 26858290; PMCID: PMC4741409.
2. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. Stat Med. 2017 Nov 30;36(27):4391-4400. doi: 10.1002/sim.7501. Epub 2017 Sep 15. PMID: 28913837; PMCID: PMC5698744.
3. Bakoyannis G, Touloumi G. Practical methods for competing risks data: a review. Stat Methods Med Res. 2012 Jun;21(3):257-72. doi: 10.1177/0962280210394479. Epub 2011 Jan 7. PMID: 21216803.
4. Collett, D. (2014). Modelling Survival Data in Medical Research (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b18041. Chapters 4,5 and 7.
5. Fine, Jason P., and Robert J. Gray. "A Proportional Hazards Model for the Subdistribution of a Competing Risk." *Journal of the American Statistical Association*, vol. 94, no. 446, 1999, pp. 496–509. *JSTOR*, https://doi.org/10.2307/2670170. Accessed 26 May 2023.
6. Ruan PK, Gray RJ. Analyses of cumulative incidence functions via non-parametric multiple imputation. Stat Med. 2008 Nov 29;27(27):5709-24. doi: 10.1002/sim.3402. PMID: 18712779.
7. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May;64(5):402-6. doi: 10.4097/kjae.2013.64.5.402. Epub 2013 May 24. PMID: 23741561; PMCID: PMC3668100.
8. Baraldi AN, Enders CK. An introduction to modern missing data analyses. J Sch Psychol. 2010 Feb;48(1):5-37. doi: 10.1016/j.jsp.2009.10.001. PMID: 20006986.
9. Bakoyannis, G., Siannis, F. and Touloumi, G. (2010), Modelling competing risks data with missing cause of failure. Statist. Med., 29: 3172-3185. https://doi.org/10.1002/sim.4133
10. LITTLE,R. J. A.RUBIN,D.B. (2002), "INTRODUCTION", " Single Imputation Methods"," Maximum Likelihood for General Patterns of Missing Data: Introduction and Theory with Ignorable Nonresponse", "BAYES AND MULTIPLE IMPUTATION" RODERICK J. A. Statistical Analysis with Missing Data Second Edition, A JOHN WILEY & SONS, INC., PUBLICATION, pp. 3-23, 59-74, 164-189, 200-222
11. Enders C.K. (2010), "An Introduction to the missing data"," Traditional Methods for Dealing with Missing Data"," Maximum Likelihood Missing Data Handling"," The Imputation Phase of Multiple Imputation", "

The Analysis and Pooling Phases of Multiple Imputation", Applied Missing Data Analysis, Todd D. Little, The Guilford Press. pp.1-36, 37 – 55, 86-126, 187-216,217-254

12. Beretta, L., Santaniello, A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inform Decis Mak* **16** (Suppl 3), 74 (2016). https://doi.org/10.1186/s12911-016-0318-z
13. Lin, WC., Tsai, CF. Missing value imputation: a review and analysis of the literature (2006–2017). *Artif Intell Rev* **53**, 1487–1509 (2020). https://doi.org/10.1007/s10462-019-09709-4
14. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011 Mar;20(1):40-9. doi: 10.1002/mpr.329. PMID: 21499542; PMCID: PMC3074241.
15. Lu K, Tsiatis AA. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. Biometrics. 2001 Dec;57(4):1191-7. doi: 10.1111/j.0006-341x.2001.01191.x. PMID: 11764260.
16. Bakoyannis, G., Zhang, Y. & Yiannoutsos, C.T. Semiparametric regression and risk prediction with competing risks data under missing cause of failure. *Lifetime Data Anal* **26**, 659–684 (2020).https://doi.org/10.1007/s10985-020-09494-1
17. Ghosh, Sujit. (1998). Bayesian Imputation Methods for Missing Data.
18. Gelman,A.,Carlin,J.B.,Stern,.H.S,Dunson,D.B.Vehtari,A.,Rubin,D.B.(2021). " Probability and inference", "Single-parameter models", "Introduction to multiparameter models", "Introduction to multiparameter models"," Asymptotics and connections to non-Bayesian approaches", " Basics of Markov chain simulation" and "Models for missing data". Bayesian Data Analysis Third edition. pp 3-27,29-57,63-79,83-98, 275-291 and 449-467.
19. Craiu RV, Duchesne T (2004) Inference based on the em algorithm for the competing risks model with masked causes of failure. Biometrika 91:543–558
20. Goetghebeur E, Ryan L (1995) Analysis of competing risks survival data when some failure types are missing. Biometrika 82:821–833
21. Hyun S, Lee J, Sun Y (2012) Proportional hazards model for competing risks data with missing cause of failure. J Stat Plan Inference 142:1767–1779
22. Albert, J. (2009)."Model Comparison". Bayesian Computation with R, second edition. Springer– Verlag, New York, pp. 181 – 201.
23. Koch,K.R.(2007). "Parameter Estimation, Confidence Regions and Hypothesis Testing", Introduction to Bayesian Statistics Second Edition. Springer, pp. 63-82

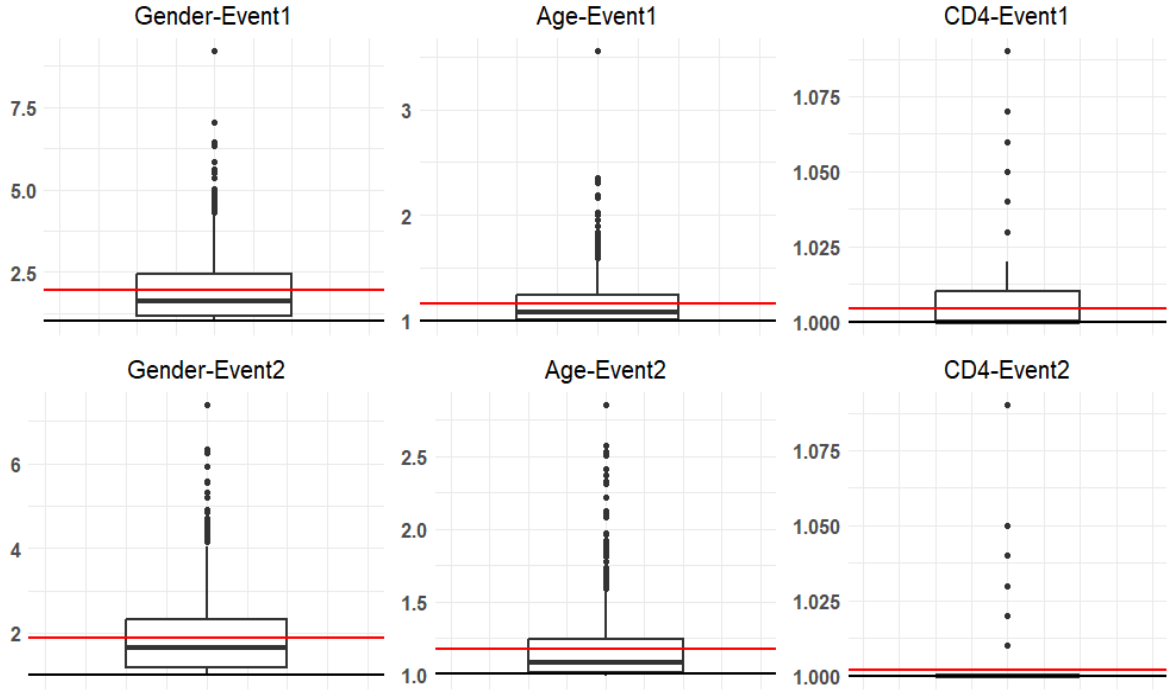
24. Robert, Christian P. Casella, George. (2009) ” Random Variable Generation”,” Monte Carlo Integration”,“Metropolis–Hastings Algorithms” and “Monitoring and Adaptation for MCMC Algorithms” Introducing Monte Carlo Methods with R, edited by Robert Gentleman, Kurt Hornik, Giovanni Parmigiani, Springer, , pp.41-60,61-88,167-195 and 237-267.
25. Ntzoufras,I.(2009), “Introduction to Generalized Linear Models: Binomial and Poisson Data “. Bayesian Modeling Using WinBUGS First Edition, A JOHN WILEY & SONS, INC., PUBLICATION, pp. 229-270.
26. Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). *bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework*. Journal of Open Source Software, 4(40), 1541. https://doi.org/10.21105/joss.01541
27. Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). *Indices of Effect Existence and Significance in the Bayesian Framework*. Frontiers in Psychology 2019;10:2767. 10.3389/fpsyg.2019.02767
28. Chris Bambey Guure, Noor Akma Ibrahim, "Bayesian Analysis of the Survival Function and Failure Rate of Weibull Distribution with Censored Data", *Mathematical Problems in Engineering*, vol. 2012, Article ID 329489, 18 pages, 2012. https://doi.org/10.1155/2012/329489
29. Sarhan, Ammar & El-Gohary, Awad & Mustafa, Abdelfattah & Tolba, Ahlam. (2020). Statistical Analysis of Regression Competing Risks Model with Covariates Using Weibull Sub-Distributions.
30. Kruschke, J.K., Liddell, T.M. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon Bull Rev* **25**, 178–206 (2018). https://doi.org/10.3758/s13423-016-1221-4
31. Allignol, A., Schumacher, M., Wanner, C. *et al.* Understanding competing risks: a simulation point of view. *BMC Med Res Methodol* **11**, 86 (2011). https://doi.org/10.1186/1471-2288-11-86
32. Mpofu P.B.(2020), STATISTICAL METHODS FOR DEALING WITH OUTCOME MISCLASSIFICATION IN STUDIES WITH COMPETING RISKS SURVIVAL OUTCOMES, Doctor of Philosophy in the Department of Biostatistics, Indiana University February 2020, page 134-135 table4.5 - Figure4.8
33. Morris, TP, White, IR, Crowther, MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*. 2019; 38: 2074– 2102. https://doi.org/10.1002/sim.8086

Appendix

Convergence

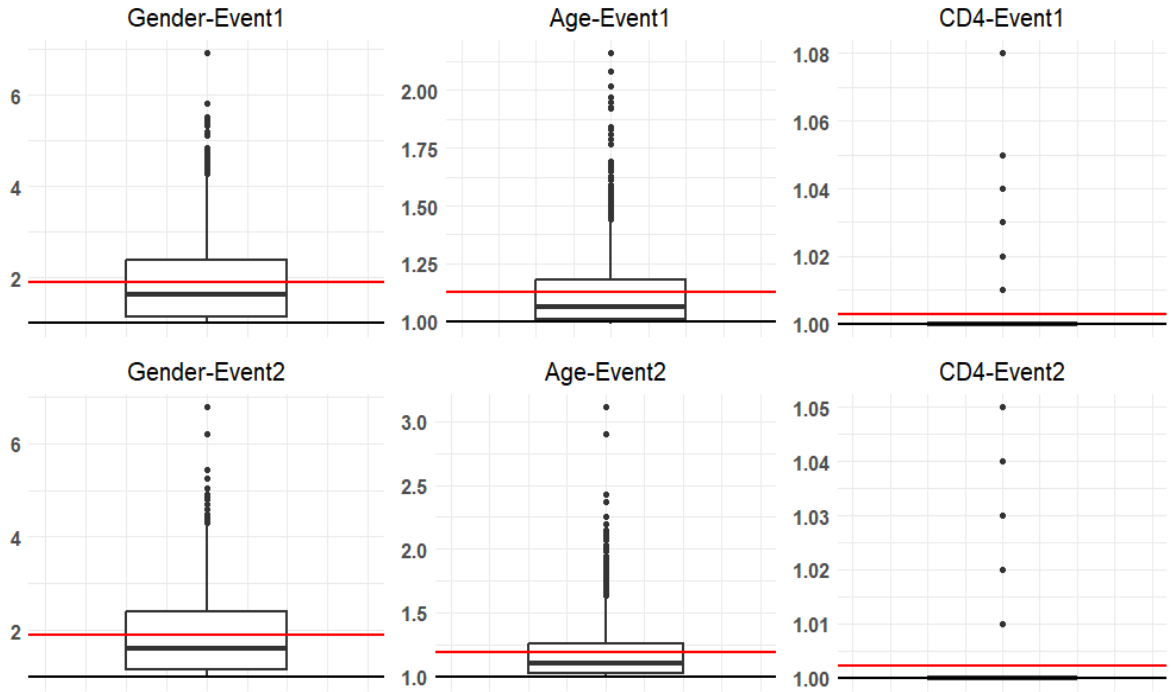
In Bayesian Simulation studies is impossible to check the convergence by graphical means such as cumulative mean plot or just plotting the sequence of simulations. Those two plots check if the mixed chains are stationary, and if they converge to a fixed distribution. This graphical method is impossible to be conducted because in practice there are 3000 datasets; therefore, for each dataset is improbable to plot those graphs. Therefore, the scale reduction factor $\hat{R}$ is used [21,24] which is referred to the second unit of this thesis [$\hat{R}$]. The scale reduction factor $\hat{R}$ converges to 1 as the number of simulations in chains tend to infinity because the perfect convergence is achieved when the number of simulations tend to infinity. When the scale reduction factor $\hat{R}$ is bigger than 1, this means that the combination of chains has not been efficiently converged because the mixed chains are not stationary, and a bigger number of simulations is needed. So, for every simulated coefficient a $\hat{R}$ has been calculated using the formula $\hat{R}$. To sum up, If the scale reduction factor is bigger than 1 the combinations of chains has been inadequately converged and more simulations are needed. In this master thesis, it is reasonable to observe scale reduction factors bigger than 1 because of the low number of chains which are two for each parameter and the low number of simulations (each chain has 3000 simulations which only 100 are chosen). For every scenario, event and parameter 1000 scale reduction factors have been estimated. As a result, plotting boxplot is a convenient way to observe and understand the distribution of scale reduction factors.

Graph 22: The distributions of Scale Reduction function $\hat{R}$ for the coefficients of scenario 1.



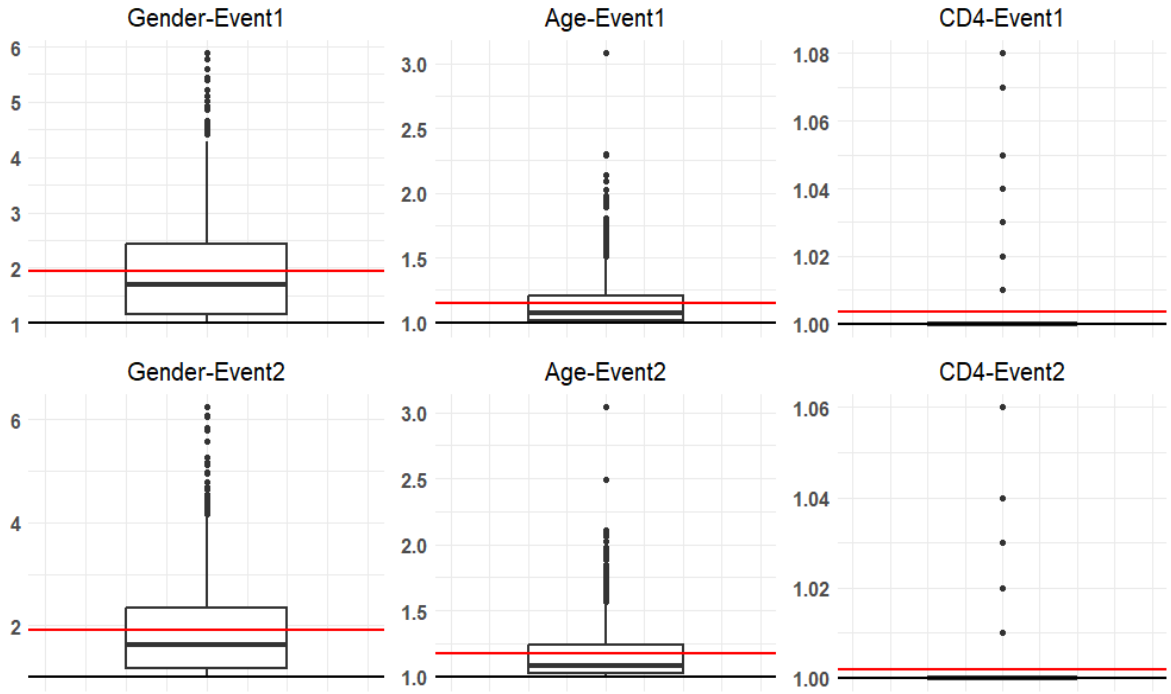
*The red line is the y=mean of $\hat{R}$ and black line in the bottom is the y=1.

Graph 23: The distributions of Scale Reduction function $\hat{R}$ for the coefficients of scenario 2



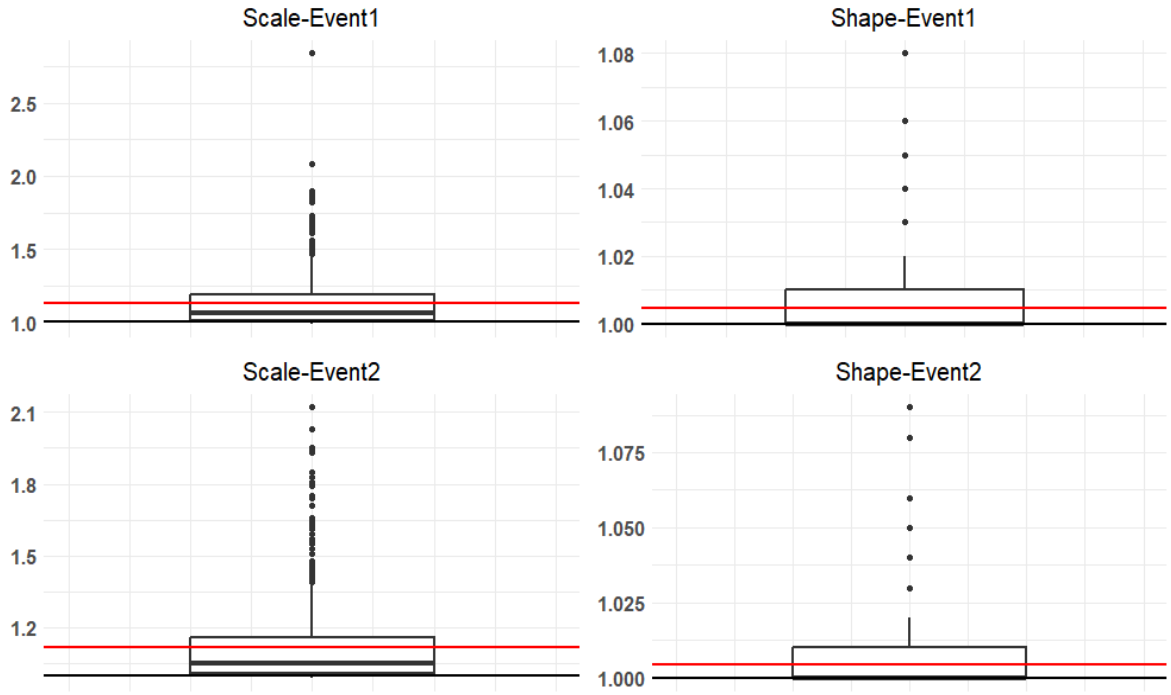
*The red line is the y=mean of $\hat{R}$ and black line in the bottom is the y=1.

Graph 24: The distributions of Scale Reduction function $\hat{R}$ for the coefficients of scenario 3.



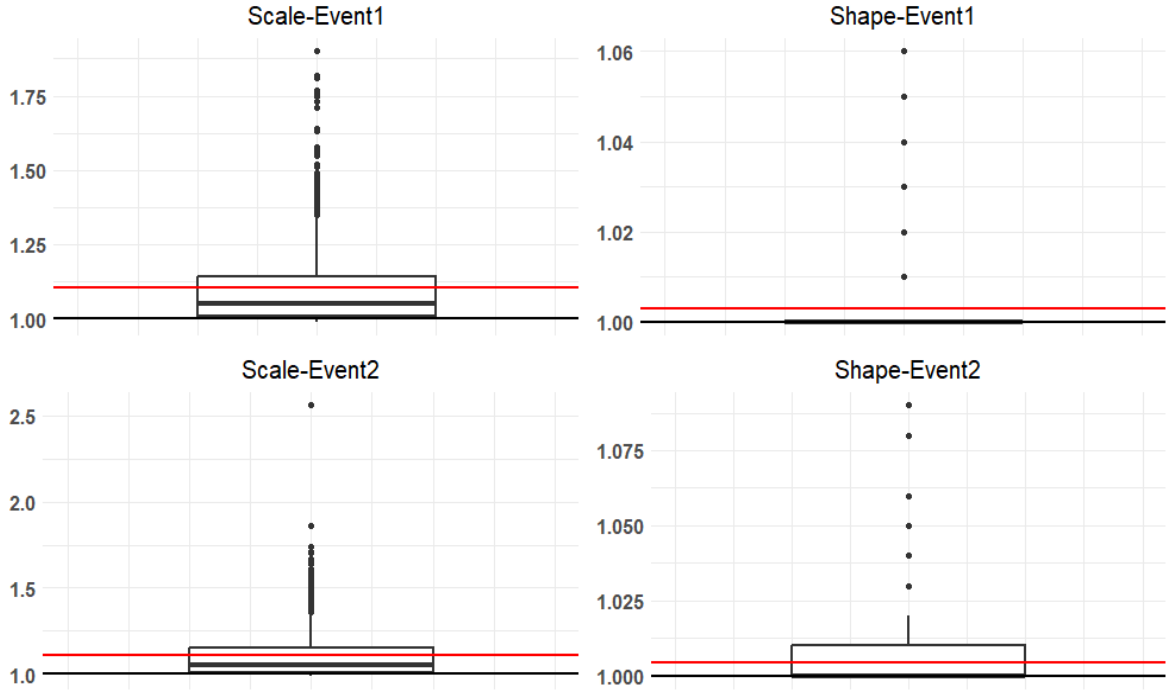
*The red line is the y=mean of $\hat{R}$ and black line in the bottom is the y=1.

Graph 25: The distributions of Scale Reduction function $\hat{R}$ for the Weibull parameters of scenario 1.



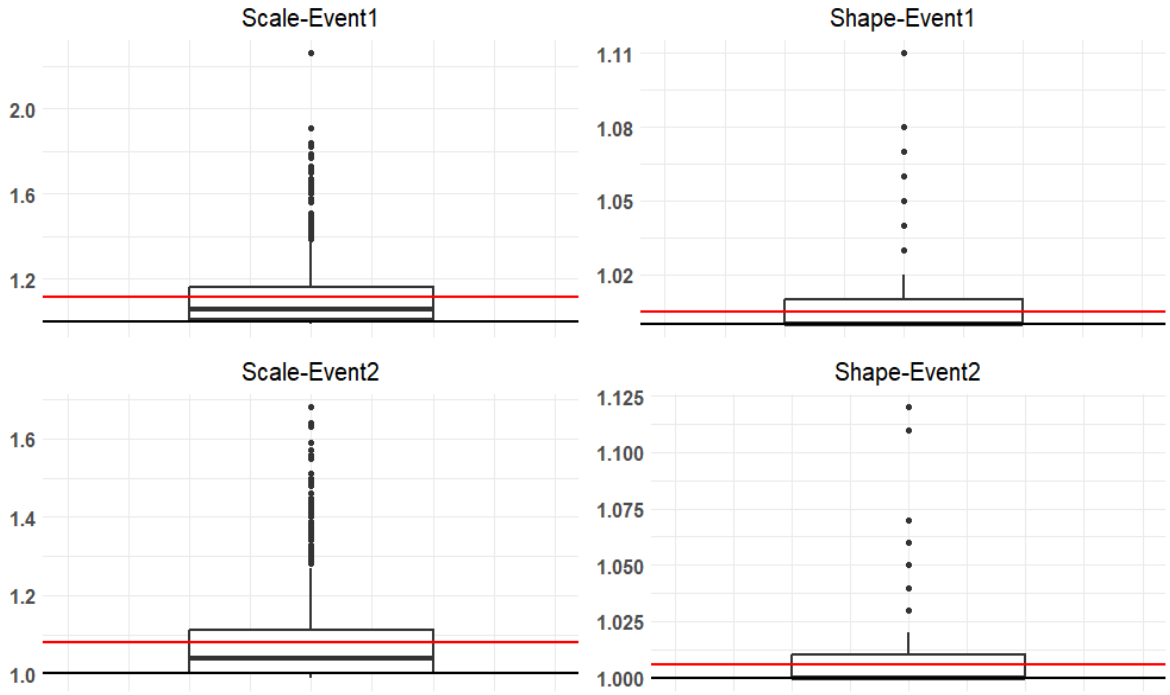
*The red line is the y=mean of $\hat{R}$ and black line in the bottom is the y=1.

Graph 26: The distributions of Scale Reduction function $\hat{R}$ for the Weibull parameters of scenario 2.



*The red line is the y=mean of $\hat{R}$ and black line in the bottom is the y=1.

Graph 27: The distributions of The Scale Reduction function $\hat{R}$ for the Weibull parameters of scenario 3.



*The red line is the y=mean of $\hat{R}$ and black line in the bottom is the y=1.

As it is observed, the mean of Gender's $\hat{R}$ is bigger than two in both two events and in all 3 scenarios. This informs us that the value of $\hat{R}$ is independent of the

scenario. Also, the mean of the scale parameter and Age's $\hat{R}$ are bigger than 1. Because of those results, the convergence of the previous quantities is uncertain. Furthermore, in some situations the $\hat{R}$ is close to 1 which indicates us that in some situations there is a convergence. Also, the results are dependent on the parameters;namely, the mean of Gender's $\hat{R}$ is pretty much the same in all events and situation. This is applied to the rest of coefficient and parameters. In addition, if the real results were unknown, a $\hat{R}$ bigger than 1 would be a concerning result. Apart from the $\hat{R}$, the convergence in this situation can be assessed by the Bias and the MSE because the real results are known. From the tables both Bias and MSE are very small, so it is rational to assume that convergence has been achieved. Because of the concerning big scale reduction factor in Gender, some further investigation is needed in order to efficiently understand the problem. It is observed that big $\hat{R}$ is a result of big range between the maximum and the minimum of the simulated coefficients and when the points are not random. In other words, if the simulations do not extremely fluctuate around a value the scale reduction factor is small. Big fluctuations cause big scale reduction factors. Basically, the optimal situation is, that all chains are stationary and all have not theoretically but practically the same distribution.Also, the chains of the simulations are not stationary ($\hat{R} > 1$) because there is a strong correlation between the values. The theory suggests increasing the number of simulations inside the chains and lowering the correlation by increasing the number of simulations between two choices (I choose one simulation every 25). Also, the prior variance of Gender is bigger than the variance of Age and CD4 ;as a result, one can say that the scale reduction function is related to the prior distribution. So, the possible solutions with the aim of eradicating the big values of $\hat{R}$ are two, one is to increase the number of chain simulations and the second one is to increase the variance of prior distribution because it is too low. The first solution is suggested from the theory [18] and the second one is an ambiguous one because it is related to just an observation (the relation of prior variance and the $\hat{R}$).

Code in R

In the Appendix, I represent the data simulation methodology and the Bayesian Weibull competing risk model with missing cause of failure. There are three scenarios, without the loss of generality, I represent only the second scenario.

Data Simulation

The distribution of the Gender, Age and CD4

```
n<-6657
probCens<-3382/n
probmissing<-2481/(n-3382)
probdeath<-445/(349+445)
```

Gender

```
FEMALE<-c(2300,210,254,1665)
MALE<-c(1082,139,191,816)
pGen<-sum(MALE)/(sum(FEMALE+MALE))
```

Age

```
meanAge<-(37.9*3382 + 35.5*349+37.3*445+35.4*2481)/n
q25Age<-(31.8*3382+29.7*349+31.3*445+29.9*2481)/n
q75Age<-(45.4*3382+41.9*349+46*445+42.7*2481)/n
sdAge<-(11+8.7)/2
```

CD4

```
meanCD4<-(174*3382+145*349+88*445+155*2481)/n
q25CD4<-(91*3382+69*349+39*445+71*2481)/n
q75CD4<-(258*3382+222*349+180*445+214*2481)/n
sdCD4<-(120+110.8)/2
```

The coefficients

```
#pcause1 DEATH
bgender1<-log(1.24)
bage1<-log(1.10)/10
bcd41<-log(0.76)/100
#pcause2 DISENGAGEMENT
bgender2<-log(1.15)
bage2<-log(0.75)/10
bcd42<-log(1.03)/100
```

The function to calculate the standard error of the coefficients

```
sdcalc<-function(lh,llower,lupper){
 sdlower<-abs((lh-llower)/1.96)
 sdupper<-abs((lh-lupper)/1.96)
 return(mean(c(sdlower,sdupper)))
}
```

The standard error of the coefficients

```
#pcause1 DEATH
sdgender1<-sdcalc(log(1.24),log(0.96),log(1.59))
sdage1<-sdcalc(log(1.10)/10,log(0.97)/10,log(1.25)/10)
sdcd41<-sdcalc(log(0.76)/100,log(0.63)/100,log(0.91)/100)
#pcause2 DISENGAGEMENT
sdgender2<-sdcalc(log(1.15),log(1.02),log(1.31))
```

```
sdage2<-sdcalc(log(0.75)/10,log(0.7)/10,log(0.8)/10)

sdcd42<-sdcalc(log(1.03)/100,log(1.0)/100,log(1.06)/100)
```

The observed distribution of the Gender, Age and CD4. This distribution is needed to assign the missingness in the different observation.

```
#Observed probability
```

Gender

```
pmaleobserved<-(139+ 191)/(349+445)
```

Age

```
AgeOBSmean<-(35.5*349 + 37.3* 445)/(349+445)

AgeOBSq25<-(29.7*349+31.3*445)/(349+445)

AgeOBSq75<-(41.9*349+46.0*445)/(349+445)

AgeOBSsd<-mean(c(8.77,11.38))
```

CD4

```
CD4OBSmean<-(145*349 + 88* 445)/(349+445)

CD4OBSq25<-(69*349+39*445)/(349+445)

CD4OBSq75<-(222*349+180*445)/(349+445)

CD4OBSsd<-mean(c(90.3,126.5))
```

The missing distribution of the variables.

```
#MISSING DISTRIBUTION
```

Age

```
AgeMISSmean<-35.4

AgeMISSsd<-mean(c(8.16,10.8))
```

Gender

```
pmalemisiing<-816/2481
```

CD4

```
CD4MISSmean<-155

CD4MISSsd<-mean(c(125,87.4))
```

The scale and shape parameters.

```
scale1<-exp(-1.2)

shape1<-0.584


shape2<-0.931

scale2<-1.028*scale1


DataLista2<-list()
```

```
for(k in 1:1000){
```

The simulation of the three variables.

```
Gender<-rbinom(n,1,pGen)

Age<-rnorm(n,meanAge,sdAge)

Age<-abs(Age)

CD4<-rnorm(n,meanCD4,sdCD4)

CD4<-abs(CD4)
```

The survival function.

```
St<-function(t,j){

 a<-exp(-
(exp(bgender1*Gender[j]+bage1*Age[j]+bcd41*CD4[j])*scale1*(t^shape1)+exp(bgender2*Gen
der[j]+bage2*Age[j]+bcd42*CD4[j])*scale2*(t^shape2)))

 return(a)

}
```

The numerical solution of the equation S(t)=u (Newton Rapson)

```
t<-numeric(n)

for(i in 1:n){

 u<-runif(1)

 t[i]<-uniroot(function(t) St(t,i)-u,
interval=c(0.00000000000000000000000000000001,300))$root

}
```

Two hazards

```
#log(p1/p2)=log(h1/h2) => p1/p2 = h1/h2 => p1 = h1/h2 /(1+h1/h2)

h1<-exp(bgender1*Gender+bage1*Age+bcd41*CD4)*scale1*shape1*(t^(shape1-1))

h2<-exp(bgender2*Gender+bage2*Age+bcd42*CD4)*scale2*shape2*(t^(shape2-1))
```

The probability of the cause of failure being death

```
pcause1<-(h1/h2)/(1+(h1/h2))

u<-runif(6657)

dobs<-1*(u<=pcause1)+2*(u>pcause1)
```

Censoring

```
tcen<-rexp(6657,rate=0.3435)

dobs<-dobs*(t<=tcen)

finalt<-apply(data.frame(t,tcen),1,min)
```

The missing and observed density

```
densmiss<-
dbinom(Gender[dobs!=0],1,pmalemisiing)*dnorm(Age[dobs!=0],AgeMISSmean,AgeMISSsd)*dn
orm(CD4[dobs!=0],CD4MISSmean,CD4MISSsd)

densobs<-
dbinom(Gender[dobs!=0],1,pmaleobserved)*dnorm(Age[dobs!=0],AgeOBSmean,AgeOBSsd)*dn
orm(CD4[dobs!=0],CD4OBSmean,CD4OBSsd)


probt<-numeric(length(finalt[dobs!=0]))
```

The time effect on the missing probability

```
probt<- 0.2*(finalt[dobs!=0]<0.1768903)+0.2*(finalt[dobs!=0]>1.5997850)+0.2

estimprobmis<-probt*(densmiss/(densmiss+densobs))

missInd<-numeric(n)

umis<-runif(length(finalt[dobs!=0]))
```

Assigning the missingness

```
missInd[dobs!=0]<-1*(umis<5.35*estimprobmis)

dobs[missInd==1]<- -1

print(k)

DataLista2[[k]]<-data.frame(dobs,finalt,missInd,Gender,Age,CD4)

}
```

The end

Bayesian Weibull competing risks with missing cause of failure

Coefficients

```
#pcause1 DEATH

bgender1<-log(1.24)

bage1<-log(1.10)/10

bcd41<-log(0.76)/100

#pcause2 DISENGAGEMENT

bgender2<-log(1.15)

bage2<-log(0.75)/10

bcd42<-log(1.03)/100
```

Scale and shape parameters

```
scale1<-exp(-1.2)

shape1<-0.584

scale2<-1.028*scale1

shape2<-0.931
```

The standard error of the coefficients

```
sdcalc<-function(lh,llower,lupper){
 sdlower<-abs((lh-llower)/1.96)
 sdupper<-abs((lh-lupper)/1.96)
 return(mean(c(sdlower,sdupper)))
}
#pcause1 DEATH
sdgender1<-sdcalc(log(1.24),log(0.96),log(1.59))
sdage1<-sdcalc(log(1.10)/10,log(0.97)/10,log(1.25)/10)
sdcd41<-sdcalc(log(0.76)/100,log(0.63)/100,log(0.91)/100)
#pcause2 DISENGAGEMENT
sdgender2<-sdcalc(log(1.15),log(1.02),log(1.31))
sdage2<-sdcalc(log(0.75)/10,log(0.7)/10,log(0.8)/10)
sdcd42<-sdcalc(log(1.03)/100,log(1.0)/100,log(1.06)/100)

library(casebase)
library(MASS)
library(emdbook)
library(car)
library(DescTools)
DataLista<-readRDS("scenario2/DataLista2.RData")

finalmatrix1<-matrix(numeric(18*1000),nrow=1000,ncol=18)
finalmatrix2<-matrix(numeric(18*1000),nrow=1000,ncol=18)
```

Scale reduction factor

```
Rhat<-function(mat){
 meann<-apply(mat,2,mean)
 varr<-apply(mat,2,var)
 W<-mean(varr)
 B<-var(meann)
 R<-round(sqrt(1- 1/nrow(mat) +B/W),2)
 return(R)
}
```

The cause of failure probability (it is used for the data imputation)

```
pcause1<-function(t,Gen,Ag,Cd){
```

```
 h1<-exp(bgender1*Gen+bage1*Ag+bcd41*Cd)*scale1*shape1*(t^(shape1-1))
 h2<-exp(bgender2*Gen+bage2*Ag+bcd42*Cd)*scale2*shape2*(t^(shape2-1))
 pc1<-(h1/h2)/(1+(h1/h2))
 return(pc1)
}
```

The essential Bayesian parameters for the first event

#competingR1=list(b1canSigma,b1mo,b1Sigma,l1canSd,l1shape,l1scale,g1canSd,g1shape,g1scale,b1=0,l1=1,g1=1)

The covariance of generator

b1canSigma=0.00000008*matrix(c(1,0,0,0,1,0,0,0,1),nrow=3, byrow=F)

The mean of the prior

b1mo=c(bgender1,bage1,bcd41)

The covariance of the prior

b1Sigma=matrix(c(sdgender1^2,0,0,0,sdage1^2,0,0,0,sdcd41^2),nrow=3, byrow=F)

The scale parameter of the first event

The standard deviation of the scale generator

l1canSd=0.013

The scale prior parameters

l1shape=scale1

l1scale=1

The shape parameter of the first event

The standard deviation of the shape generator

g1canSd=0.019

The shape prior parameters

g1shape=shape1

g1scale=1

The initial values

b1=c(bgender1,bage1,bcd41)

l1=scale1

g1=shape1

competingR1=list(b1canSigma,b1mo,b1Sigma,l1canSd,l1shape,l1scale,g1canSd,g1shape,g1scale,b1=b1,l1,g1)

The essential Bayesian parameters for the second event

#competingR2=list(b2canSigma,b2mo,b2Sigma,l2canSd,l2mo,l2Sd,g2canSd,g2shape,g2scale,b2=0,l2=1,g2=1)

The covariance of generator

b2canSigma=0.00000007*matrix(c(1,0,0,0,1,0,0,0,1),nrow=3, byrow=T)

The mean of the prior

b2mo=c(bgender2,bage2,bcd42)

The covariance of the prior

b2Sigma=matrix(c(sdgender2^2,0,0,0,sdage2^2,0,0,0,sdcd42^2),nrow=3, byrow=T)

The scale parameter of the second event

The standard deviation of the scale generator

l2canSd=0.018

The scale prior parameters

l2shape=scale2

l2scale=1

The shape parameter of the second event

The standard deviation of the shape generator

g2canSd=0.038

The shape prior parameters

g2shape=shape2

g2scale=1

The initial values

b2=c(bgender2,bage2,bcd42)

l2=scale2

g2=shape2

competingR2=list(b2canSigma,b2mo,b2Sigma,l2canSd,l2shape,l2scale,g2canSd,g2shape,g2scale,
b2=b2,l2,g2)

Metropolis Hasting

#nround is the number of the simulations

MHvalue<-
function(nround,fdata,competingR1=list(b1canSigma,b1mo,b1Sigma,l1canSd,l1shape,l1scale,g1
canSd,g1shape,g1scale,b1=0,l1=1,g1=1),competingR2=list(b2canSigma,b2mo,b2Sigma,l2canSd,
l2shape,l2scale,g2canSd,g2shape,g2scale,b2=0,l2=1,g2=1)){

lx<-3

simulationmatrixFIRST<-matrix(numeric((lx+2)*nround),nrow = nround,ncol=lx+2)

simulationmatrixSECOND<-matrix(numeric((lx+2)*nround),nrow = nround,ncol=lx+2)

```
pb1<-numeric(nround)

pl1<-numeric(nround)

pg1<-numeric(nround)

pb2<-numeric(nround)

pl2<-numeric(nround)

pg2<-numeric(nround)
```

#Initial coefficients of the predictive model

```
glm1<-
glm(dobs==1~log(finalt)+Gender+Age+CD4,data=fdata[fdata$missInd==FALSE&fdata$dobs
%in%c(1,2),],family = binomial(link = "logit"))

theta<-glm1$coefficients

thetaMat<-matrix(numeric(5*nround),nrow=nround,ncol=5)

for(i in 1:nround){
```

#Imputation Step

```
 #pc<-
pcause1(fdata$finalt[fdata$missInd==1],fdata$Gender[fdata$missInd==1],fdata$Age[fdata$mis
sInd==1],fdata$CD4[fdata$missInd==1])

 pc<-
gcause1(fdata$finalt[fdata$missInd==1],fdata$Gender[fdata$missInd==1],fdata$Age[fdata$mis
sInd==1],fdata$CD4[fdata$missInd==1],theta)

 u<-runif(sum(fdata$missInd==1))

 fdata$dobs[fdata$missInd==1]<-1*(u<=pc)+2*(u>pc)
```

#Update step

#Thelo kai ta imputed mazi

```
 glm1<-
glm(dobs==1~log(finalt)+Gender+Age+CD4,data=fdata[fdata$dobs%in%c(1,2),],family =
binomial(link = "logit"))

 theta<-mvrnorm(1,glm1$coefficients,vcov(glm1))


 thetaMat[i,]<-theta


 t<-fdata$finalt

 dobs<-fdata$dobs

 x<-matrix(c(fdata$Gender,fdata$Age,fdata$CD4),nrow = nrow(fdata),ncol = 3,byrow = F
```

First event

```
 dth<-1

 rate<-0
```

The generation of the first event coefficients

```
################ b1  ###############
b1can<-mvrnorm(1,b1,b1canSigma)

expcan<-exp(x%*%b1can)
expj<-exp(x%*%b1)
```

Two likelihoods

```
likeCAN<-(expcan^(dobs==dth))*exp(-expcan*(l1*(t^g1)))
likej<-(expj^(dobs==dth))*exp(-expj*(l1*(t^g1)))
```

The acceptance probability

```
rate<-prod((likeCAN/likej))*(dmvnorm(b1can,b1mo,b1Sigma)/dmvnorm(b1,b1mo,b1Sigma))
rate
if (is.na(rate)|is.nan(rate)){rate<- 0}

pb1[i]<-min(1,rate)
u<-runif(1)
simulationmatrixFIRST[i,1:lx]<-b1can*(u<=pb1[i])+b1*(u>pb1[i])
b1<-simulationmatrixFIRST[i,1:lx]

###
expj<-exp(x%*%b1)
##
```

The scale parameter simulation

```
################ l1 ###############
l1can<-rnorm(1,l1,l1canSd)
rate<-0
if(l1can>0){

likeCAN<-((l1can)^(dobs==dth))*exp(-expj*(l1can*(t^g1)))
likej<-((l1)^(dobs==dth))*exp(-expj*(l1*(t^g1)))
rate<-
prod((likeCAN/likej))*(dgamma(l1can,shape=l1shape,scale=l1scale)/dgamma(l1,shape=l1sha
pe,scale=l1scale))

if (is.na(rate)|is.nan(rate)){rate<-0}
```

```
pl1[i]<-min(1,rate)
u<-runif(1)
simulationmatrixFIRST[i,lx+1]<-l1can*(u<=pl1[i])+l1*(u>pl1[i])
l1<-simulationmatrixFIRST[i,lx+1]
} else {
 simulationmatrixFIRST[i,lx+1]<-l1
 pl1[i]<-0
}
```

The shape parameter simulation

```
################ g1 ###################
g1can<-rnorm(1,g1,g1canSd)
rate<-0
if(g1can>0){

 likeCAN<-((g1can*(t^(g1can-1)))^(dobs==dth))*exp(-expj*(l1*(t^g1can)))
 likej<-((g1*(t^(g1-1)))^(dobs==dth))*exp(-expj*(l1*(t^g1)))



 rate<-
prod((likeCAN/likej))*(dgamma(g1can,shape=g1shape,scale=g1scale)/dgamma(g1,shape=g1s
hape,scale=g1scale))



 if (is.na(rate)|is.nan(rate)){rate<-0}


 pg1[i]<-min(1,rate)
 u<-runif(1)
 simulationmatrixFIRST[i,lx+2]<-g1can*(u<=pg1[i])+g1*(u>pg1[i])
 g1<-simulationmatrixFIRST[i,lx+2]
} else {
 simulationmatrixFIRST[i,lx+2]<-g1
 pg1[i]<-0
}


p1matrix<-matrix(c(pb1,pl1,pg1),ncol=3, byrow=F)
```

Second event

#SECOND EVENT

The generation of the second event coefficients

```
dth<-2
rate<-0
################ b2  ###############
b2can<-mvrnorm(1,b2,b2canSigma)

expcan<-exp(x%*%b2can)
expj<-exp(x%*%b2)
```

Two likelihoods

```
likeCAN<-(expcan^(dobs==dth))*exp(-expcan*(l2*(t^g2)))
likej<-(expj^(dobs==dth))*exp(-expj*(l2*(t^g2)))
```

The acceptance probability

```
rate<-prod((likeCAN/likej))*(dmvnorm(b2can,b2mo,b2Sigma)/dmvnorm(b2,b2mo,b2Sigma))

if (is.na(rate)|is.nan(rate)){rate<- 0}

pb2[i]<-min(1,rate)
u<-runif(1)
simulationmatrixSECOND[i,1:lx]<-b2can*(u<=pb2[i])+b2*(u>pb2[i])
b2<-simulationmatrixSECOND[i,1:lx]

###
expj<-exp(x%*%b2)
##
```

The scale parameter simulation

```
###############  l2  ###############
l2can<-rnorm(1,l2,l2canSd)
rate<-0
if(l2can>0){

 likeCAN<-((l2can)^(dobs==dth))*exp(-expj*(l2can*(t^g2)))
 likej<-((l2)^(dobs==dth))*exp(-expj*(l2*(t^g2)))
```

```
 rate<-
prod((likeCAN/likej))*(dgamma(l2can,shape=l2shape,scale=l2scale)/dgamma(l2,shape=l2sha
pe,scale=l2scale))


 if (is.na(rate)|is.nan(rate)){rate<-0}


 pl2[i]<-min(1,rate)
 u<-runif(1)
 simulationmatrixSECOND[i,lx+1]<-l2can*(u<=pl2[i])+l2*(u>pl2[i])
 l2<-simulationmatrixSECOND[i,lx+1]
} else {
 simulationmatrixSECOND[i,lx+1]<-l2
 pl2[i]<-0
}
```

The shape parameter simulation

```
################ g2 ###################


g2can<-rnorm(1,g2,g2canSd)
rate<-0
if(g2can>0){


 likeCAN<-((g2can*(t^(g2can-1)))^(dobs==dth))*exp(-expj*(l2*(t^g2can)))
 likej<-((g2*(t^(g2-1)))^(dobs==dth))*exp(-expj*(l2*(t^g2)))


 rate<-
(prod(likeCAN/likej))*(dgamma(g2can,shape=g2shape,scale=g2scale)/dgamma(g2,shape=g2s
hape,scale=g2scale))


 if (is.na(rate)|is.nan(rate)){rate<-0}


 pg2[i]<-min(1,rate)
 u<-runif(1)
 simulationmatrixSECOND[i,lx+2]<-g2can*(u<=pg2[i])+g2*(u>pg2[i])
 g2<-simulationmatrixSECOND[i,lx+2]
} else {
 simulationmatrixSECOND[i,lx+2]<-g2
```

```
 pg2[i]<-0
}




}

p1matrix<-matrix(c(pb1,pl1,pg1),ncol=3, byrow=F)

p2matrix<-matrix(c(pb2,pl2,pg2),ncol=3, byrow=F)

return(list(simulations1=simulationmatrixFIRST,simulations2=simulationmatrixSECOND,p1mean=apply(p1matrix,2,mean),p2mean=apply(p2matrix,2,mean)))
}
```

I divided the simulation to three sessions. The first begins from the first dataset and ends to the 333th dataset. The second session begins from 334 and ends to 667. The third one starts from 668 and ends to 1000

```
for(k in 1:333){
 fdata<-DataLista[[k]]
```

2 chains. Each chain does 3000 simulations.

```
mh1<-MHvalue(3000,fdata,competingR1,competingR2)
 mh2<-MHvalue(3000,fdata,competingR1,competingR2)
```

500 simulations are discarded and after that one final simulation pair 25 is chosen

```
 sim11 <- mh1$simulations1[501:3000,]
 sim11<-sim11[seq(1,nrow(sim11),by=25),]
 sim21<-mh1$simulations2[501:3000,]
 sim21<-sim21[seq(1,nrow(sim21),by=25),]



 sim12 <- mh2$simulations1[501:3000,]
 sim12<-sim12[seq(1,nrow(sim12),by=25),]
 sim22<-mh2$simulations2[501:3000,]
 sim22<-sim22[seq(1,nrow(sim22),by=25),]
```

```
 sim1<-rbind(sim11,sim12)
 sim2<-rbind(sim21,sim22)

 finalmatrix1[k,1:5]<-apply(sim1,2,mean)
 finalmatrix1[k,6:10]<-apply(sim1,2,sd)
 finalmatrix1[k,11:13]<-apply(rbind(mh1$p1mean,mh2$p1mean),2,mean)

 finalmatrix2[k,1:5]<-apply(sim2,2,mean)
 finalmatrix2[k,6:10]<-apply(sim2,2,sd)
 finalmatrix2[k,11:13]<-apply(rbind(mh1$p2mean,mh2$p2mean),2,mean)
```

The calculation of scale reduction factor

```
for (j in 1:ncol(sim1)){
  mat<-matrix(sim1[,j],ncol=2,byrow = F)
  finalmatrix1[k,13+j]<-Rhat(mat=mat)
 }
 for (j in 1:ncol(sim2)){
  mat<-matrix(sim2[,j],ncol=2,byrow = F)
  finalmatrix2[k,13+j]<-Rhat(mat=mat)
 }
 theta1<-mh1$thetaSim[501:3000,]
 theta2<-mh2$thetaSim[501:3000,]

 theta1<-theta1[seq(1,nrow(theta1),by=25),]
 theta2<-theta2[seq(1,nrow(theta2),by=25),]

 finalTheta[k,]<-apply(rbind(theta1,theta2),2,mean)
 print(k)
}
Finalmatrix1
finalmatrix2
finalTheta
```

The end