POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

A neural network approach to survival analysis with time-dependent covariates for modelling time to Cardiovascular diseases in HIV patients

Tesi di Laurea Magistrale in
Mathematical Engineering - Statistical Learning

Author: **Agostino Lurani Cernuschi**

Student ID: 943027
Advisor: Prof. Chiara Masci
Co-advisors: Ing. Federica Corso
Academic Year: 2020-21

Abstract

In this thesis we analyse the impact of AntiRetroviral Therapies (ARTs) drugs and other clinical measurements on the time to a CardioVascular Diseases (CVDs) event in HIV patients through the application of survival methods to a real dataset. We consider two different approaches: the classical Proportional Hazard (PH) Cox model and a neural network based method, the DeepHit. First, we analyse and compare the two methods fitted on clinical and therapies data measured at the baseline, i.e., at the beginning of the ART. Then, we move to a time-dependent setting, considering the whole follow-up of patients, and we analyse and compare extensions of the two methods: the time-dependent Cox PH model and the Dynamic DeepHit. All methods are compared in terms of interpretability and predictive performances. The compared models have similar performances and are able to reach high values of Concordance-index (0.77). The neural network method is more flexible than the Cox model, it relaxes the PH assumption and allows to capture non-linear and time-varying relationships between the covariates and the target variable. On the other side, it has two weaknesses: the computational cost, that becomes prohibitive with time-dependent data, and the difficulty of interpretation. This last problem is turned into a point of strength with the use of the Shapley Additive Explanation that enabled to interpret and visualise the interaction between covariates and the time to CVD events. These models provide very interesting results: short time of exposure to ART inhibitor drugs increases the risk of CVD events in 15 years, while long time of exposure to these drugs is a strong protective factor.

Keywords: Survival Analysis, Neural Network, DeepHit, Time-Dependent Data, HIV, Cardiovascular Disease.

Abstract in lingua italiana

In questa tesi, analizziamo l'impatto che specifici medicinali usati nelle Terapie AntiRetrovirali (ART) e le caratteristiche cliniche e personali di pazienti affetti da HIV hanno sul rischio di un evento cardiovascolare, applicando metodi di analisi di sopravvivenza su dati reali. Consideriamo due approcci differenti: il classico modello Cox Proportional Hazard (PH) e il DeepHit, un metodo basato su reti neurali. Prima, analizziamo e confrontiamo i due metodi considerando i dati clinici e della terapia alla baseline, i.e. i dati misurati all'inizio della ART. In seguito, includiamo dati dipendenti dal tempo, misurati sullo storico delle visite dei pazienti, e applichiamo e confrontiamo due altri metodi: il modello di Cox tempo-dipendente e il Dynamic DeepHit. I modelli vengono valutati e confrontati in termini di interpretabilità e capacità predittiva. I due approcci mostrano performance predittive simili e raggiungono valori alti di C-index (0.77). Il metodo basato su reti neurali permette di rilassare l'ipotesi di PH e di linearità degli effetti delle covariate, risultando più flessibile del modello di Cox. Dall'altro lato, presenta due debolezze: il costo computazionale, che diventa proibitivo con i dati tempo dipendenti, e la difficoltà di interpretazione dei risultati. L'applicazione delle tecniche di interpretazione degli Shapley values, che permettono di interpretare e visualizzare l'interazione tra due variabili e il tempo all'evento cardiovascolare, trasforma la seconda debolezza in un punto di forza. I modelli applicati producono interessanti risultati: una breve esposizione ai medicinali delle ART aumenta di poco il rischio di un evento cardiovascolare a 15 anni, mentre una lunga esposizione diminuisce fortemente questo rischio.

Parole chiave: Survival Analysis, Neural Network, DeepHit, Dati Tempo-Dipendenti, HIV, Malattie Cardiovascolari.

Contents

Introduction

Across the world over 37 millions of people are affected by Human Immunodeficiency Virus [1] (HIV). This virus attacks and destroys CD4 cells, which are cells that fight against infections. A Loss of CD4 cells makes the immune system more vulnerable to diseases and infections and if not treated, it can lead to Acquired Immunodeficiency Syndrome (AIDS). The World Health Organization, WHO, declares that *"HIV continues to be a major global public health issue"* [2]. The most critical issue is that there is no cure for HIV infection. In the late 90s the introduction of the AntiRetroiral Therapy (ART), has enabled people infected with HIV to lead longer lives.
The ART is based on the administration of different types of drugs, each belonging to a particular group of inhibitors based on the stage of the cycle they inhibit: the Nucleotide Reverse Transcriptase Inhibitors (NRTIs), the Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs) and the Protease Inhibitors (PIs). In 1996 the life expectancy of a 20-year-old man affected by HIV was of 19 years [26]. In 1995 the protease inhibitors were introduced and only after 1998 [29] the ART has been developed in a combination of the three groups of drugs, resulting in a drop of deaths of almost a half [22]. Despite the fact that life expectancy has lengthened, HIV infection is nowadays associated with CardioVascular Diseases [7] (CVDs) and it is not clear though if this event is a side effect of ARTs. Many studies have been conducted about the relationship between these drugs and the CVD risk, but the results are not conclusive. Each category of inhibitors includes many drugs that have different relationships with CVD events. There is evidence from observational studies that the use of PIs increases risk of Myocardial Infarction [31] (MI). *"There exists an increased risk of myocardial infarction in patients exposed to Abacavir and Didanosine within the preceding 6 months"* and *"the risk of MI was increased by cumulative exposure to all the studied PIs except saquinavir"* [11] . The other two inhibitors' categories do not seem to increase the risk of CVD as much as PIs. There is no evidence of increasing risk after the exposure of any NRTIs except abacavir and there are no associations between MI and the exposure to NRTIs [17]. In a systematic review of risk of CVD events from ART for HIV [5], Clay Bavinger et all conclude that *"based on the overall evidence, we believe there is still uncertainty whether ART leads to increased*

cardiovascular risk, and if so, the magnitude of that risk". Even though the relationships between each class of inhibitors and CVD are partially unclear, it is noted that some inhibitors increase the risk of CVD depending on the timing of the exposure. These studies were conducted over patients with short follow-up, e.g. in [17] patients have been followed for 6 years. It is worth it to specify that these studies do not consider ARTs that have started after 2007, year in which the first integrase inhibitor (INI) was approved. This class of inhibitor drug offers a different way to prevent HIV from making copies of itself.

From a statistical point of view, recent studies explore both the risk of CVD events and the time to CVD in HIV patients. Some studies apply classification methods to identify patients characteristics associated to the risk of CVD, such as machine-learning methods [24]. Others analysed the time to event with regression methods, such as Lasso regression [20] or adaptive boosting, naive Bayes, K-nearest neighbours or random forest [8]. The most sophisticated methods take into account and aim to model the time to CVD event, if any, with survival analysis approaches and to identify features associated with it [9] [25], and some of them have used time-dependent data [28]. The most-widely used method to estimate the time to event analysis is the Cox PH (PH) model [6]. This method estimates the probability curves of CVD event for each patient and adjust them for a set of covariates. The Cox PH model makes strong assumptions on the distribution of the covariates and on their linear association to the target variable. Moreover it assumes the effects of the covariates to be time-invariant. Other models, such as the conditional logistic regression [17] or the Poisson regression [31], assume different form of relationships between covariates and parameters but they do not remove the assumption of the relationships being time-invariant.

The infectious diseases department of the IRCCS San Raffaele have collected the clinical history of patients affected by HIV from 1998 until now. The long duration over time of the follow-ups and the fact that there are present patients that have started the ART before and after 2007 allow to study the long-term effects of the ARTs on CVDs and also to analyse how the introduction of INIs has changed these effects. The dataset consists in the ARTs inhibitor drugs measurements of 4512 individuals that as well as general and clinical information. In this work we analyse the time to the CVD event in HIV patients within 15 years from the beginning of the ART and the relationship between ART drugs and time to CVD events applying survival analysis tools on the dataset supplied by IRCCS San Raffaele. Firstly, we investigate this relationship considering only data at the baseline, i.e. clinical and personal information at the beginning of the ART, as time independent covariates. Then we include the tracking of these information through time, considering time-dependent covariates that regard time to exposure to drugs, the time of a diagnosis

and the variation of some clinical measurement across time. Given the complexity of the dataset and the possible unknown interaction among covariates, we compare the Cox PH model with a new neural network method, the DeepHit [18], to overcome the needs of weakening the assumption of time-independency between covariates and time-to-event and to be able to analyse also non-linear relationships. It is noted that recently new approaches based on neural network methods have been proposed [10], however, these studies do not have weakened the assumption of the time-dependency between covariates and the time-to-event. In particular, we compare the classic Cox PH model [6] and the DeepHit in terms of performances and complexities. When considering longitudinal data, i.e. the follow up of each patient, in order to capture how the relationship between ART and CVD varies over time, we apply time-dependent Cox PH model and the Dynamic DeepHit model [19], defined as the extensions of the previous models for time-dependent data.

Therefore the goal of this work is twofold:

i) To compare neural networks approach with classic survival analysis methods in a complex framework such as the survival analysis, initially with models at the baseline and then including time-dependent covariates.

ii) To capture the relationship between covariates and the risk of cardiovascular disease, on a long term time horizon of 15 years, of a person affected by HIV, with a special focus to ART inhibitors.

The thesis is organized as follows. First we introduce the statistical methods, starting with the introduction to survival analysis tools: basics knowledge and the Kaplan-Meier estimators [15] (chapter 1), and the classic Cox model and its extended version (Chapter 2). Than, after a brief guide to neural networks, the structure of DeepHit (Chapter 3), and Dynamic DeepHit (Chapter 4), is explained. After the introduction of the dataset used (Chapter 5), we will compare the baseline methods in terms of interpretability, complexity and performance (Chapter 7). Then we will perform the two models with longitudinal data (Chapter 8), where we also study the time dependency in the relationship between ART and CVD. We will analyse the behaviour of previous models on a bootstrapped dataset (Chapter 9) and draw the conclutions (Chapter 10).

All the analysis is performed using R [23] and Python [30].

1 | Survival analysis

1.1. Introduction

Survival analysis is the statistical process that studies the time until an event occurs, i.e. the period that passes from a starting point until an event occurs. The time to the event is commonly called *survival time* to describe an individual that has "survived" over the follow up period. We will also refer to it also as *time-to-event*. Survival analysis is commonly used in clinical fields but it can be adopted to describe the lifetime of an object such as the functioning time of a battery. Although the majority of its application intends the event as a failure, e.g. death or disease incidence, time-to-event may be a positive experience such the recovering time of an illness. The time-to-event can refer to a single event or more than one, in this case we talk about competing risks.

1.1.1. Censoring

The main problem to deal with, when approaching survival analysis, concerns censoring data. We can have individuals that leave the study before the event of interest happens or we may know that the event has happened a certain time before but do not know precisely when. We usually refer to these individuals, for whom the event does not occur during the follow-up, as *censored* data. In particular, two types of censoring can be considered and treated in different ways:

- right censoring: occurs when a person does not experience the event before the end of the study or we lost the information about the patient's health during the follow-up. In this case we do not know when the event of interest will happen, but we know that the patient was alive until the censoring moment. Therefore the true time-to-event could be greater or equal to the observed one.
- left censoring: occurs when we can not observe the exact time when the event occurred. Therefore, the real time-to-event could be shorter or equal to the observed one.

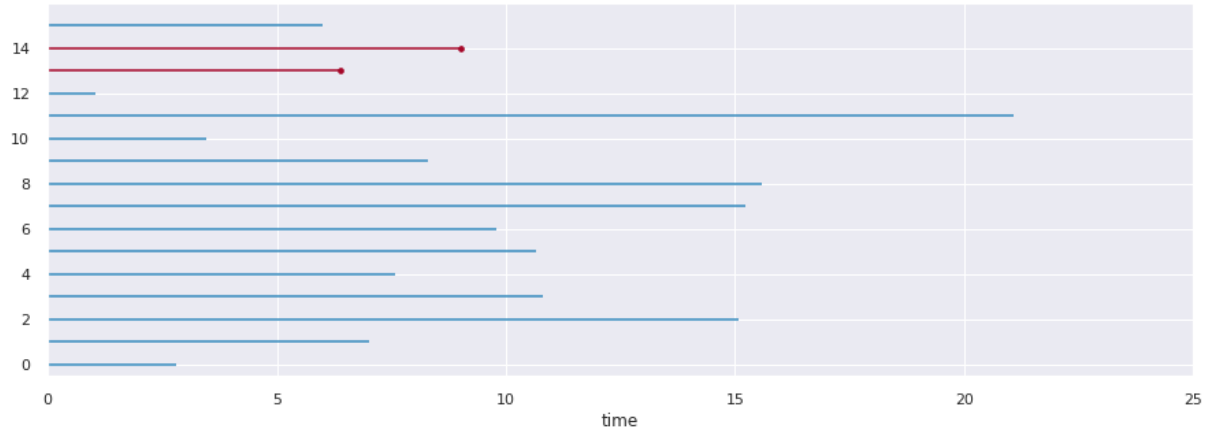

Figure 1.1: Survival time of 16 patients. Time is represented on the x-axis, while censored patients are shown in blue and those who experienced the event in red.

In Figure 1.1 a sample of sixteen patients and their time-to-event is shown. Survival time is represented on the x-axis, censored patients are shown in blue while those who experienced an event are in red.

1.1.2. Survival and hazard functions

In survival analysis theory the variable of interest is the minimum between the the time-to-event and the time of censoring as follows:

$$T = min(T^*, C) \tag{1.1}$$

where:

- T^* is a non negative random variable representing the true time to event.
- C is a non negative random variable representing the time of censoring.
- T is the observed time.

The survival time is now represented by the couple of the two random variables: the observed time and the indicator of censored data.

$$D = \{(T_i, \delta_i), i = 1, ..., N\} \tag{1.2}$$

where:

- $\delta_i = \mathbb{I}(T_i^* \leq C_i)$ is the indicator random variable that indicates whether an i^{th} observation is censored or not.
- N is the total number of observations.

The *survival function* $S(t)$ is introduced to represent the probability of survival at each time t of an individual. Considering the random variable of survival time T, its density $f(t)$ and distribution $F(t) = P(t \leq T)$, we can define the survival function as:

$$S(t) = P(T > t) = 1 - \mathbb{P}(T \leq t) = 1 - F(t) \tag{1.3}$$

From this definition we can introduce the concept of instantaneous risk of failure which is described by the *hazard function*:

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \tag{1.4}$$

The relationship between the two functions is expressed as follows and it will be important also in the definition of the KM estimator:

$$h(t) = h(t_j) = \mathbb{P}(T = t_j | T \geq t_j) = \frac{f(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})} \tag{1.5}$$

where:

- $t_1 \leq t_2 \leq ... \leq t_J$ is the ordered series of the time of all the events.
- t_j is the time of event of individual j, j $\in$ {1:J} where J is the total number of events.

A first estimator that we may use for the survival function is the proportion of survived individuals over the total number of individuals. This function measures an empirical probability of surviving at each time t:

$$\widehat{S}(t) = \frac{s_t}{N} \tag{1.6}$$

where:

- s_t is the number of *survived* individuals at time t.
- N is the total number of individuals.

Here we have the problem of handling censored patients since we know the percentage of patients that did not survive until time t_j, but we do not know whether patients censored before t_{j-1} are going to experience the event before the time t_j or not.

1.2. KM estimator

In this section, in order to handle the problem of censored data, we introduce a non-parametric statistic used to estimate the survival function $S(t)$. The Kaplan-Meier estimator [15] (KM) is defined as the probability of surviving in a certain time interval considering the time random variable as discretized in many small intervals.

First, we have to make three strong assumptions that are:

- censoring is independent from survival time: $T \perp C$.
- the survival probabilities are the same for patients recruited in different periods of the study.
- the events occurred at the specified times, i.e. we assume that there are no left censored data.

The KM estimator is defined as:

$$\widehat{S}(t) = \prod_{j:t_j^* \leq t} p_j = \prod_{j:t_j^* \leq t} (1 - \frac{d_j}{n_j}) \tag{1.7}$$

where:

- t_j^* is the time of manifestation of the event, j $\in$ {1:J} where J is the total number of events.
- p_j is the conditional probability of survival time t_j^*
- d_j is the number of observed events during Δt: $\mathbb{P}(T > t_j^* | T \geq t_{j-1}^*)$.
- n_j is the number of survived individuals before time t_j^*.

The KM estimator may also be derived from the likelihood of the hazard function, which is:

$$\mathcal{L}(t) = \prod_{j=1}^{J} P(T_j = t_j^*)^{d_j} \cdot P(T_j > t_j^*)^{n_j - d_j} = \prod_{j=1}^{J} h_j^{d_j} \cdot (1 - h_j)^{n_j - d_j} \tag{1.8}$$

The maximum likelihood estimator of h_j measured by the minimization of the negative log-likelihood is: $\hat{h}_j = \frac{d_j}{n_j}$ from which we get to the KM estimator $\widehat{S}(t)$.

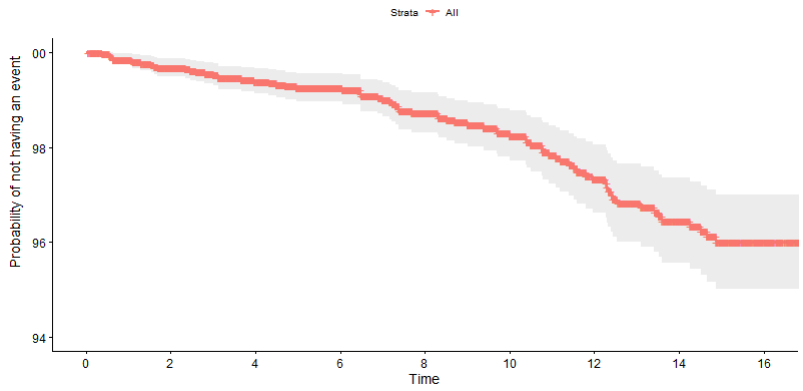

Figure 1.2: KM estimator curve for the probability of CVD.

In Figure 1.2 the KM estimator curve is represented with its 95% confidence interval. We can note that the curve is not continuous, moreover at time 0 it has a value of 100% and decreases until the time is equal to 15 years, almost 96%, after which the curve remains constant because no more events are observed over that time. The 95% confidence intervals are also shown and their width depends on the number of individuals that did not have the event until that time.

1.3. Log-rank test

The log-rank ratio test is a methodology for comparing the distributions of survival times in two or more different and independent groups. For instance in this work, we may want to verify whether patients affected by hepatitis C have a lower probability of CVD event until a certain point rather than those who are not. In this case we can divide the patients in three groups ($HCV+$, $HCV-$ and HCV_{NK}) and analyze the two survival functions setting the null hypothesis as:

$$H_0 : S_{HCV_+}(t) = S_{HCV_-}(t) = S_{HCV_{NK}}(t) \tag{1.9}$$

and the alternative as H_1: at least one of the followings is true:

$$\begin{aligned} S_{HCV_+}(t) &\neq S_{HCV_-}(t), \\ S_{HCV_-}(t) &\neq S_{HCV_{NK}}(t), \\ S_{HCV_+}(t) &\neq S_{HCV_{NK}}(t). \end{aligned} \tag{1.10}$$

We can use the KM estimator for each different group to estimate and represent the

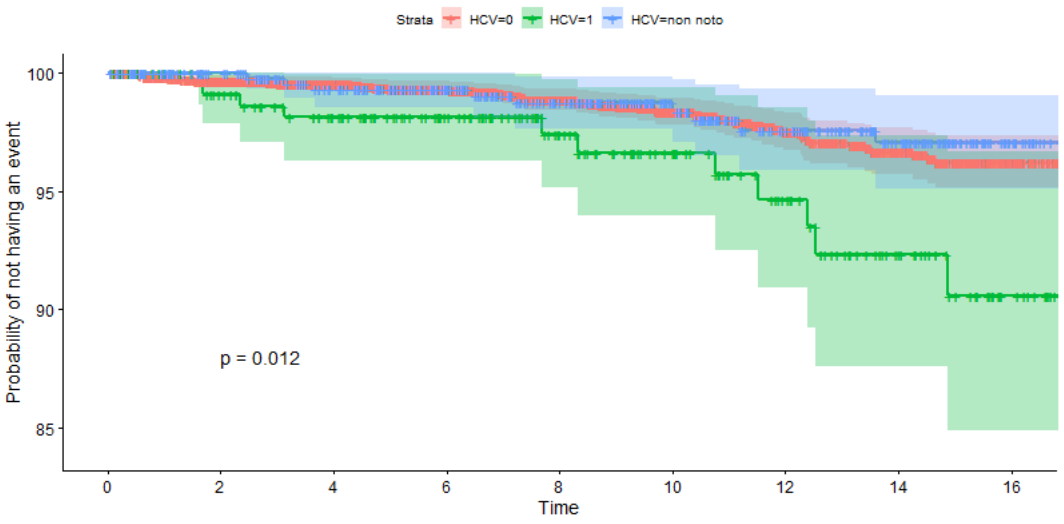

Figure 1.3: KM estimator curves for hepatitis C. (1: HCV positive; 0: HCV negative; not-known: HCV not known).

survival functions.
In Figure 1.3 the three KM estimator curves are shown. The green curve is significantly different from the others and it suggests that the group of patients with hepatitis C has a lower probability of not having the cardiovascular disease.

The basic idea for testing the null hypothesis is to compare the observed events and the expected ones in each of the following intervals $[t^*_{j-i}; t^*_j)$. First we compute the probability of an event to happen at time t^*_j, then we consider the expected number of events as the product between this probability and the number of patients at risk at time t^*_j:

$$e_{kj} = p_{dj} n_{kj} = \frac{d_j}{n_j} n_{kj} \tag{1.11}$$

where

- k represents the belonging group.
- J is the total number of events.
- n_{kj} is the number of patients at risk in group k at time t^*_j.
- d_j is the number of events in the interval $[t^*_{j-i}; t^*_j)$.

We can sum up the number of observed patients and the expected ones as follows:

$$O_k = \sum_{j=1}^{J} d_{kj} \qquad E_k = \sum_{j=1}^{J} e_{kj} \tag{1.12}$$

Finally we use a $Chi - squared$ test to investigate H_0 with the following statistic:

$$\chi^2 = \sum_{k=1,\dots,3} \frac{(O_k - E_k)^2}{E_k} \tag{1.13}$$

Log-rank test				
	N	Observed	Expected	χ^2 statistic
HCV-	3598	63	69.6	2.802
HCV+	292	15	5.9	14.031
HCV Not-known	622	12	0.419	0.501
Chisq. 12.1 on 2 d.o.f. with p-value = 5e-4				

Table 1.1: Log-rank test table for hepatitis C example with the frequency of patients in each group, the observed and the expected numbers of those who have the event within each group, the $Chi - squared$ statistics and the p-value.

In Table 1.1 we have evidence to reject H_0 since the p-value is $5{\cdot}10^{-4}$ and we can argue that the KM estimator curves between patients with and without hepatitis C are significantly different.

1.4. Hazard Ratio

The log-rank test tells us whether two or more KM estimator curves are statistically different but it does not quantify how much a group is more likely to survive than another. We can measure the differences between the two groups by using the so called Hazard Ratio (HR). It is defined as the ratio between the probability of survival in group 1 over the probability of survival in group 2 (e.g., treated vs control group). Under the hypothesis that the hazard ratio is constant over time we can test if the ratio is equal, smaller or bigger than one. It quantifies the strength of association of a dependent variable with the event of interest. The hazard ratio is usually presented with a confidence interval, which represents the reliable range of values in which we expect the true value of the hazard ratio to be included. The hazard ratio is defined as:

$$HR = \frac{O_1/E_1}{O_2/E_2} \tag{1.14}$$

We can interpret the result as:

- HR $= 1$, there is no difference;
- HR < 1, group 1 is a protective factor, i.e. patients in group 1 are less likely to have an event;
- HR > 1, group 1 is a risk factor, i.e. patients in group 1 are more likely to have an event.

It is worth to mention that all the previous techniques are univariate, i.e. we study the effect of a single variable at the time. In the following chapter we will introduce the multivariate models for survival analysis.

2 | Semi-parametric regression models for survival analysis

In this chapter we will introduce the Cox PH Model [6] which is a semi-parametric regression model that examines the relationship of independent variables with time-to-event.

2.1. Cox PH model

The Cox proportional hazards model is a multivariate regression model introduced by Cox in 1972 [6]. Here we are going to present its main features: the maximum likelihood estimation of the parameters, the definition of the hazard ratio and the PH assumption (PH).

2.1.1. Hazard ratio

First we consider the covariate-dependent survival function $S_{\mathbb{X}}(t)$:

$$S_i(t) = S_0(t)^{\mathbb{X}_i^T \beta} \tag{2.1}$$

where:

- $S_0(t)$ is the baseline survival function.
- ${X_i}^T$ is the covariates' vector of the i^{th} individual.
- β is the vector of dimension p+1, p being the number of covariates of regression coefficients we are going to estimate.

If we define the formula in terms of hazards, given $h_0(t)$ as the baseline hazard function, we obtain at time t for the i^{th} patient the following expression:

$$h_i(t|\mathbb{X}_i) = h_0(t)e^{\mathbb{X}_i^T \beta} \tag{2.2}$$

In particular $h_0(t)$ is the unspecified function that makes the Cox PH model semi-parametric. A key feature of the Cox model is the Proportional Hazard (PH) assumption according to which the baseline hazard function $h_0(t)$ depends only on t and it does not involve $\mathbb{X}_i$. In contrast, the exponential form of $h_i(t)$ involves $\mathbb{X}_i$, but does not involve time t, meaning that $\mathbb{X}_i$ is time-independent. Since the PH assumption assumes that the hazard is constant over time, we can equivalently say that the baseline hazard for an individual is proportional to the hazard for any other individual. This proportionality is constant and independent of time. The vector β of the coefficients is estimated through the maximization of the partial likelihood. Before doing so we introduce the hazard ratio for a covariate l, which is defined as $HR_l = exp(\beta_l)$, and it gives us the possibility of interpretation of the model. Indeed if the hazard ratio of the covariate l is smaller than one it means that this covariate is protective, i.e. a increment of such covariate leads to a decrease in the probability of CVD events. One can actually make the same argumentation with the coefficients β_l with respect to zero.

Note that the Cox model implies that the hazard ratio between two patients with covariates $\mathbb{X}_i$ and $\mathbb{X}_k$, which are fixed, is constant over time:

$$HR = \frac{h_i(t|\mathbb{X}_i)}{h_k(t|\mathbb{X}_k)} = \frac{h_0(t)e^{\mathbb{X}_i\beta}}{h_0(t)e^{\mathbb{X}_k\beta}} = e^{\{(\mathbb{X}_i - \mathbb{X}_k)\beta\}} \tag{2.3}$$

We can deduce from equation 2.3 that the hazard function of the i^{th} patient is proportional to the hazard of the k^{th} patient. Moreover this relationship is constant over time and the proportion is equal to $e^{\{(\mathbb{X}_i - \mathbb{X}_k)\beta\}}$.

2.1.2. Likelihood

Typically, the formulation of a likelihood function is based on the outcome distribution. Instead, in the Cox model we do not assume a distribution for the response variable since the Cox likelihood is based on the observed order of the events rather than their joint distribution. Thus the Cox likelihood is called a "partial" likelihood. The partial likelihood considers only the probabilities for all patients that experience the event and not for the censored ones. The probability of having the event for the i^{th} individual at time t_j^* , given that she/he is still at risk until time t_{j-1}^* is:

$$L_j = \frac{e^{\mathbb{X}_j^T}}{\sum_{k \in R(t_j^*)} e^{\mathbb{X}_k^T \beta}} \tag{2.4}$$

We can notice that L_j is independent from $h_0(t)$ and so from time. Now, if we take the product of L_j over all the patients who died during the study, we obtain:

$$L(\beta) = \prod_{j=1}^{J} L_j = \prod_{j=1}^{J} \frac{e^{\mathbb{X}_j^T \beta}}{\sum_{k \in R(t_j^*)} e^{\mathbb{X}_k^T \beta}} \tag{2.5}$$

To maximize the likelihood we can transform it in logarithmic scale getting:

$$\mathbf{l}(\beta) = ln(L(\beta)) = \sum_{j=1}^{J} \left[\mathbb{X}_j^T \beta - ln \left(\sum_{k \in R(t_j^*)} e^{\mathbb{X}_k^T \beta} \right) \right] \tag{2.6}$$

where:

- $R(t_j^*)$ is the set of patients at risk at the time t_j^*.

And finally we obtain the vector of coefficients $\hat{\beta}$ as:

$$\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^p} l(\beta) \tag{2.7}$$

Moreover we can observe the baseline hazard is cancelled out in each term. Thus, the baseline hazard does not need to be specified in the Cox model, since it does not play a role in the estimation of the regression coefficients.

2.2. Cox model for time dependent variables

In some cases we want to consider covariates $\mathbb{X}$ that vary over time. Such $\mathbb{X}(t)$ are called time-dependent variables. If we consider them as longitudinal data we can still use the Cox model but, since it no longer satisfies the PH assumption, we need to use an extended version [14].

2.2.1. Hazard ratio

When time-dependent and time-independent predictor variables are considered, we can write the time-dependent Cox model that incorporates both types as shown below. The formula 2.8 reflects the same structure of the time-independent hazard ratio with a baseline hazard function $h_0(t)$ multiplied by an exponential factor.

$$h_i(t|\mathbb{X}(t)) = h_0(t)e^{\left[\sum_{k=1}^{P_{fix}} \mathbb{X}_{ik}\beta_k + \sum_{z=1}^{P_{td}} \mathbb{X}_{iz}(t)\delta_z\right]} \tag{2.8}$$

where:

- $\mathbb{X}_{ik}(t)$ are the k^{th} covariates of the i^{th} patient, both the time-dependent and independent.
- P_{fix} is the number of fixed variables.
- P_{td} is the number of time-dependent variables.
- β is the vector of coefficient of the fixed variables.
- δ is the vector of coefficient of the time-dependent variables.

An important assumption of this extended version is that the hazard at time $t$depends only on the value of the time-dependent covariates at the same time t. Moreover, the model provides only one coefficient for each time-dependent variable (i.e., δ does not depend on time). Thus, this coefficient represents the total effect of the corresponding time-dependent variable, considering all the times at which the variable was measured.

So unlike a Cox PH model, the model with time-dependent covariates computes the risk of event at each time interval and re-evaluates at which risk group a individual belongs according to whether she/he experienced the event by that time.

2.2.2. Likelihood

Similarly to the Cox model, the coefficients of time-dependent variables are estimated using a maximum likelihood approach. In this case the partial likelihood function becomes as follows:

$$L(\beta, \delta) = \prod_{j=1}^{J} L_j = \prod_{j=1}^{J} \frac{exp(\mathbb{X}_{fix_j}{}^T\beta + \mathbb{X}_{td_j}{}^T(t)\delta)}{\sum_{k \in R(t_j^*)} exp(exp(\mathbb{X}_{fix_j}{}^T\beta + \mathbb{X}_{td_j}{}^T(t)\delta)} \tag{2.9}$$

where:

- P is the number of events.
- $\mathbb{X}_{fix_j}{}^T\beta$ is the product between the fixed variables and the corresponding coefficients of patient j.
- $\mathbb{X}_{td_j}^T(t)delta$ is the product between the time-dependent variables and the corresponding coefficients.

- $R(t_j^*)$ is the set of patients at risk a time t_j^*.

The most important difference with respect to the simpler Cox model is that the proportional hazards assumption is no longer satisfied. The formula for the hazard ratio between the i^{th} and the k^{th} individuals that derives from the time-dependent Cox model is:

$$\widehat{HR} = \frac{h_i(t|\beta,\delta)}{h_k(t|,\delta)} = exp\left[\sum_{p=1}^{P_{fix}}(\mathbb{X}_{ip}^T - \mathbb{X}_{kp}^T)\beta_p + \sum_{p=P_{fix}+1}^{P_{td}}(\mathbb{X}_{ip}^T(t) - \mathbb{X}_{kp}^T(t))\delta_p\right] \tag{2.10}$$

where:

- $h_i(t|\beta,\delta)$ is the hazard function given the two vectors of coefficients of the i^{th} patient.
- $\mathbb{X}_{ipj}(t)^T\delta_p$ is the product between the time-dependent variable p of the i^{th} patient and the corresponding coefficient δ_p.
- $\mathbb{X}_{kpj}(t)^T\beta_p$ is the product between the time-independent variable p of the i^{th} patient and the corresponding coefficient β_p.

3 | DeepHit

In this section, we introduce a completely new approach to survival analysis based on machine learning techniques. The advantage of using a machine learning approach over a Cox model consists in the possibility of relaxing the hypotheses underlying the Cox model and the ability of the model of capturing non linear relationships between covariate and outcome. On the other hand, the main drawbacks connected to *black box* models are the loss of interpretability (e.g., for the Cox model by means of the hazard ratio) and the huge computational cost. At the same time, we have to outline that big steps were made in the direction of making black boxes more explainable. We show in the next sections some of these methods. In this work, we applied the DeepHit model, a neural network method for survival analysis presented in *DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks* [18] by C. Lee et al., as an alternative approach to the classical Cox model.

3.1. Neural networks

A brief introduction to neural networks is necessary to understand the method used in this research. This methodology was invented around 1969 but only in recent years, with the advent of great computational power, it has been widely used. Neural networks are very versatile and are in fact used to classify images or videos, to interpret and translate text, to handle big data regressions and to build survival models.

3.1.1. Architecture

Neural networks try to emulate a human brain building a net of synapses and neurons. A basic feed-forward neural network is built up of different layers of neurons, each connected to the following by synapses which are mathematically represented by parameters. This mathematical representation of our brain tries to replicate the ability of learning from its own mistakes. In the same way as our neurons pass information via synapses to nearby neurons, the neurons of the net use parameters to communicate among each others. A synapse which is used frequently it becomes important, in mathematical terms by it gaining higher weight. So when many observations share a common feature they reinforce the same pathway, attracting, in future, similar data.

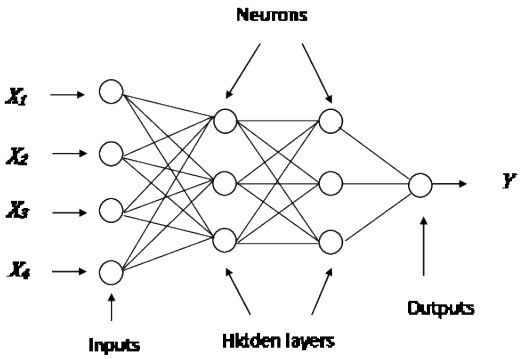

Figure 3.1: Basic neural network structure with fully-connected layers.

The net is structured by layers, an input layer that takes data as input, some hidden layers of neurons that process data and the final output layer that predicts the output. Hidden layers characterize the type of network. For regression and classification problems, as well as survival analysis problems, they are called fully-connected layers. In Figure 3.1 the structure of a network with fully-connected layers is reported. Each layer is a vector of parameters, and that they are different from convolutional layers, used in image classification, where layers are matrices of parameters. The structure of the net depends also on the size of each layer, i.e. the number of neurons or parameters, and on the number of hidden layers. So the number of parameters to be estimated will be proportional to the complexity of the architecture.

3.1.2. Optimization

Neural networks are able to capture non linear and even non proportional relations. To do so they need a complex architecture therefore they have a high number of parameters even if it is not necessarily true that a bigger network performs better than a small one. The process of estimating these parameters is made by the minimization of a loss function that is used to minimize the error (e.g. the mean squared error in regression problems). A common choice is to adopt the negative maximum likelihood, in which the model makes predictions in order to match the data distribution of the target variable.
In general, with so many parameters to tune or methods to try, it is important to be able to train models fast. The gradient descent is a numeric method largely used in optimization fields but for neural networks it is computationally not suitable because of the size of the parameters. The first numerical process proposed, able to solve this problem, is the stochastic gradient descent which is an iterative method that approximates the gradient of the loss function measuring it on a random subset. Nowadays new methods such as mini-batch stochastic gradient descent, which measures the gradient in many small random subsets, called batches, make the process much faster. This modified version of stochastic gradient descent makes an iteration, i.e. updates the parameters, through the optimization of the loss function locally in a single batch. The period

during which the process visits all the batches is called *epoch* and it is equivalent to a single iteration with the stochastic gradient descent. The most used method is the Adam optimizer [13], short for Adaptive Moment Estimation, which is an adaptive method based on a momentum that exponentially smooths the gradient speeding up the process. It works in small batches as well as the mini-batch stochastic gradient descent.

3.1.3. The overfitting problem and possible solutions

The optimization process includes the selection of hyper-parameters such as the number of hidden layers and the number of neurons for each layer. Tuning these parameters could be a difficult and long process since a single training of the net can take a lot of time, even with great computational power. Furthermore, in optimization problems, the algorithm convergence to the optimal minimum could be difficult. For instance a large neural network may be able to learn perfectly the training data structure but with the risk of overfitting and of a high computational cost, while a small network could not be able to perform a good prediction. As already said a big net may cause overfitting, indeed it rapidly interpolates training data and underperforms on new ones. The first solution to prevent overfitting is to use a validation set to evaluate the model on new data and use this performance to decide when to stop the training of the model.

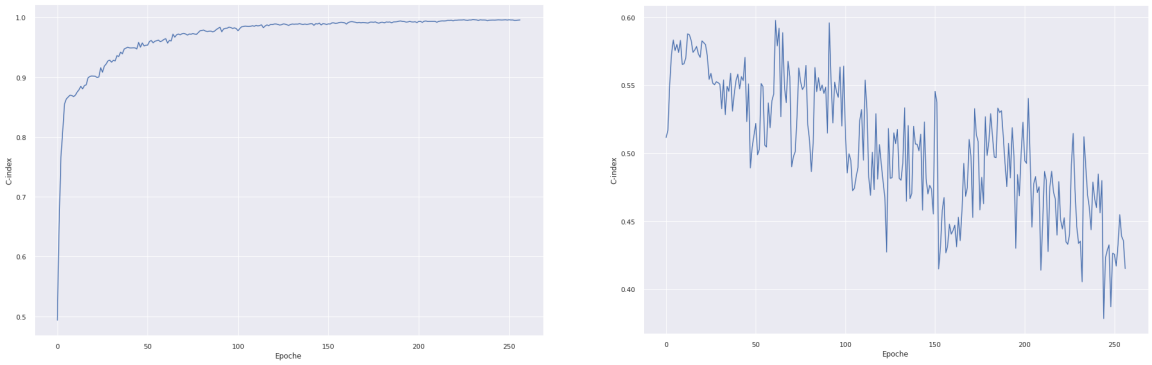

Figure 3.2: Learning curves on training and test sets. The epochs are represented on the x-axis while the concordance index (C-index) is reported on the y-axis.

This process is called *early stopping* and in Figure 3.2 are represented the learning curves of the same model on training set, on the left, and validation set, on the right. On the left the training curve reaches quickly great values and after a few epochs goes to 1. On the right figure the test curve is more variable and after some oscillations starts to decrease. These two behaviours suggest that the model is interpolating training data and it will not be a good and robust predictor. Early stopping stops the process when the validation loss, in this case the C-index, i.e. the metric used in survival analysis, starts decreasing.

If the size of the sample does not allow to split the data into three subsets: training, validation and test, and a cross-validation approach is unfeasible due to high computationality cost, two other methods may be use to prevent overfitting. In this work two other methods have been used:

- *Weight regularization* constrains parameters weight to a maximum value, similarly to the Lasso and Ridge methods.
- *Dropout* sets a random percentage of the parameters weights to zero.

Both methods need a tuning process for the selection of the two hyper-parameters: the maximum magnitude and the percentage of dropout.

3.2. DeepHit

Before entering into the details of DeepHit [18] it is important to underline that the goal is to estimate the joint distribution of the first hitting time, i.e. the time-to-event. The major pros of this method with respect to the Cox model are the relaxation of the PH assumption and that it allows both the parameters and the time to event random variable to depend on the covariates.

3.2.1. Architecture

The architecture of the DeepHit net is slightly different from standard ones, having k cause-specific networks, in order to handle multiple events, and one that is shared between all possible competing risks. See Figure 3.3. The shared sub-network takes as inputs all the covariates $\mathbb{X}$ and produces as output a function, $f_s(\mathbb{X})$, that captures the representation common to the patients that experiment an event. The other k sub-networks are cause-specific meaning that they capture the features common to patients of the k^{th} competing risk. An important strength of this model is therefore the ability to handle competing risks, i.e. it can handle also situations in which there are multiple causes for an event. In this work the interest has been posed on a single event, grouping all cardiovascular diseases together.
By using DeepHit we do not have to make any assumption on the underlying stochastic process, so that both, the parameters and the form of the stochastic process, could depend on the covariates.
Each cause-specific sub-network takes as inputs the pairs of full covariates and the resulting function of the shared sub-network: $Z = (f_s(\mathbb{X}), \mathbb{X})$. It produces as output a vector $f_{c_k}(Z)$, which corresponds to the probability of event of a specific cause k. It is important to underline that the k sub-networks take as input also the initial covariates that are not being processed by the shared network. This gives the sub-networks access to the learned common representation $f_s(\mathbb{X})$ but also allows them to learn the non-common part of the representation.

The last layer, the one that produces the output, is a unique Softmax layer for all sub-networks. The Softmax function produces the joint distribution of competing risk and it returns the probability of belonging to each class. It is defined as:

$$f(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \tag{3.1}$$

Where:

- z_i is the marginal probability of belonging to the class i = 1,..., K.
- K is the total number of possible classes

In this way the output of the total network is not the marginal distribution for each cause-event but it is the joint distribution, $y = [y_{1,1},, y_{1,T_{max}},, y_{K,1},, y_{K,T_{max}}]$. A patient with covariates $\mathbb{X}$, has an output element $y_{k,s}$ which is the estimate of the probability, $\widehat{P}(s, k|\mathbb{X})$, that the patient will experience the event k at time s. In such a way the network is able to capture the relationships between covariates and risks, which can be non-linear and also non-proportional.

The cause-specific cumulative incidence function expresses the probability that a particular event $k^* \in K$ occurs on or before time t^* conditional on the covariates $\mathbb{X}^*$. This represents the key of the survival analysis under competing risks. By definition, the cumulative incidence function for the event k^* is:

$$F_{k^*}(t^*|\mathbb{X}^*) = \mathbb{P}(s \leq t^*, k = k^*|\mathbb{X} = \mathbb{X}^*) = \sum_{s^*=0}^{t^*} \mathbb{P}(s^* = t^*, k = k^*|\mathbb{X} = \mathbb{X}^*) \tag{3.2}$$

However, since the true cumulative incidence function, $F_{k^*}(s^*|\mathbb{X}^*)$, is not known, we utilize the estimate, $\widehat{F}_{k^*}(s^*|\mathbb{X}^*) = \sum_{m=0}^{s^*} y_{k,m}^*$ in order to compare the risk of an event occurring and to assess how models discriminate across cause-specific risks among patients.

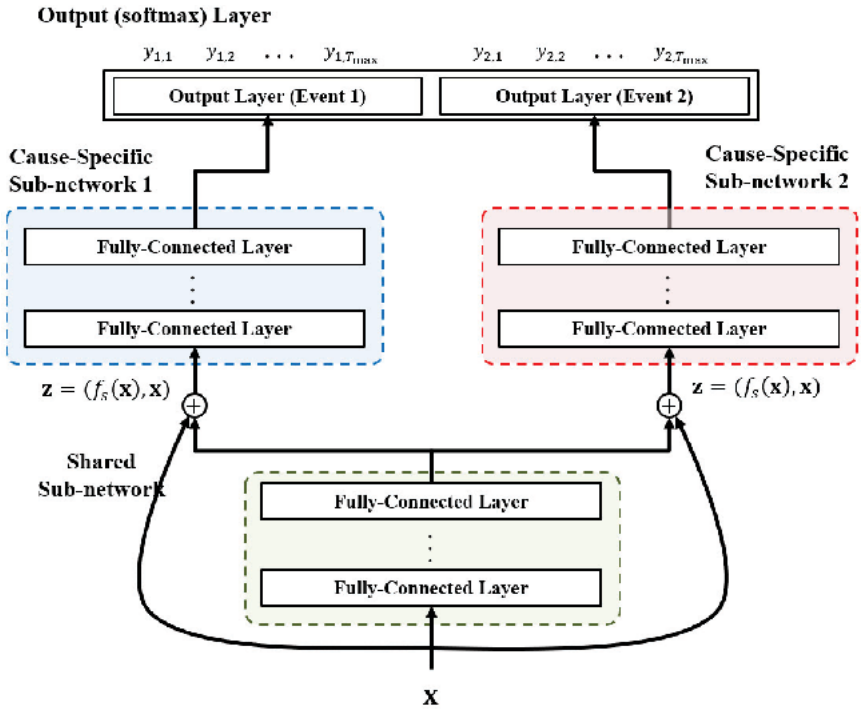

Figure 3.3: The architecture of DeepHit. The two curves at the bottom directly connect X to the k-cause specific sub-network.

3.2.2. Loss function

Another important feature that characterizes DeepHit is the *loss function* which is a function of the parameters to be estimated. To minimize this function we are going to use an approximation with the mini-batch stochastic gradient descent to make the process quicker. The loss function used in DeepHit is the sum of two terms: $L_{Total} = \alpha L_1 + \beta L_2$.
L_1 is the log-likelihood of the joint distribution of the first hitting time. L_2 incorporates estimated cumulative incidence functions calculated at different times (i.e. the time at which an event actually occurs) in order to finetune the network to each cause-specific estimated cumulative incidence function.

The log-likelihood function has to be modified to take into account right censored data. In particular for those patients who are not censored we want to capture the exact time of event while for censored patients, we want the time of lost in follow-up, which only tells us that such a patient was alive until that moment. This is represented by the loss function with the first term regarding uncensored data and the second term regarding those who are lost in follow-up:

$$L_1 = -\sum_{i=1}^{N}[\mathbb{I}(k^i \neq \emptyset) log(y^i_{k^i,s^i})] + \mathbb{I}(k^i = \emptyset) log(1 - \sum_{k=1}^{K} \widehat{F}_k(s^i|X^i)) \tag{3.3}$$

The second loss function, L_2, tries to make the model able to predict lower probability of event for patients who have had the event earlier. It can be interpreted as the C-index. In fact, we can say that our model is performing well if, comparing two patients, the one with lower hitting time is predicted with a lower probability of event. We represent this as $A_{k,i,j} = \mathbb{I}(k^i = k, s^i = s^j)$ and define the concordance loss function as:

$$L_2 = \sum_{k=1}^{K} \alpha_k \sum_{i \neq j} (A_{k,i,j} \eta(\widehat{F}_k(s^i|\mathbb{X}^i), \widehat{F}_k(s^j|\mathbb{X}^j)) \tag{3.4}$$

where:

- $\eta(x, y) = exp(\frac{x-y}{\sigma})$.
- $\alpha_k = 1, \forall k = 1...K$.
- $\widehat{F}_k(s^j|\mathbb{X}^j)$ is the estimated cumulative function of the specific cause k for the j^{th} individual at the time s^j.

4 | Dynamic DeepHit

In this section we introduce a new approach to handle time-dependent data and capture their relationship with survival time over the observation period. The machine learning model adopted in this section is an extended version of DeepHit for time-dependent covariates introduced in 2020 by C. Lee et al. [19].

4.1. Recurrent neural networks

Handling dynamic variables, i.e. time dependent, requires the introduction of a slightly different kind of net: recurrent neural network [32]. Feed-forward neural networks, are based on processing data points that are independent one from each other and so are unfeasible for data series or any kind of auto-correlated data.

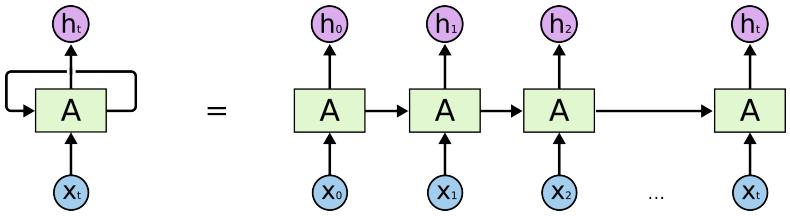

Figure 4.1: Feedback loop and unrolled layer, the input $\mathbb{X}_t$ enters into the layer A then the output h_t goes both into the next layer and back into its own layer.

In order to deal with time-dependent data it is necessary to capture the auto-correlation of data series. The idea of recurrent neural networks is to give some kind of memory to each layer enabling it to process also old data points. The practical solution consists in a feedback loop on each layer, so that the input of each layer consists not only in the output of the previous one and also in its own output. Furthermore, it is possible to feedback the outputs of distant layers enabling the network to catch all previous information. On the right side of Figure 4.1 there is an unrolled representation of the same process, here the figure shows that the block A at time t^* has as input all previous $\mathbb{X}_t$ with $t \leq t^*$. This explains how the memory of recurrent layers enables the network to capture time-dependent features.

4.2. Architecture

Dynamic DeepHit has a similar architecture to the one with time independent variables. Indeed it preserves the first shared block, a family of cause-specific sub-networks and at the end a Sofmax function as output layer.

The important difference with respect to the previous model is that the shared block is composed of a recurrent neural network and an attention mechanism, that helps to find more important the parts of previous time-dependent measurements when making predictions. The output of the shared block is the only input of each sub-network, since sub-networks do not handle directly original data as was in the DeepHit model, but they capture the latent patterns that characterize different competing events.

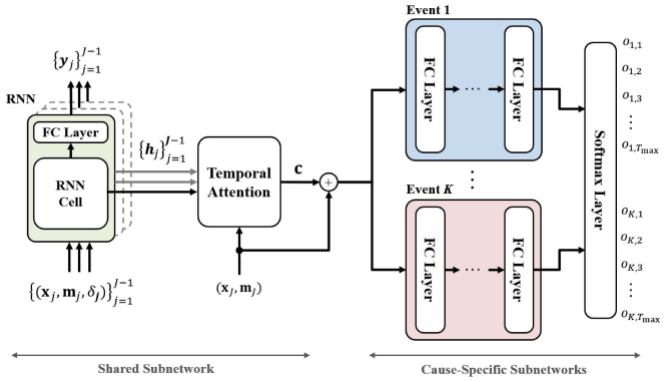

Figure 4.2: Architecture of Dynamic DeepHit. The first part is the shared sub-network that capture the time dependent features common to any competing event. Every cause specific sub-network captures the feature of one particular risk.

Dynamic-DeepHit uses a Softmax layer in order to summarize the outcomes of each cause-specific sub-network, $f_{c_1}(), ..., f_{c_K}()$, and to map them into a proper probability measure. Overall, the network produces an estimated joint distribution of the first hitting time and of the competing events. In particular, given a subject with $\mathbb{X}^*$, each output node represents the probability of having k^{th} event at time t, i.e., $o^*_{k,t} = \widehat{P}(T = t, k = k|\mathbb{X}^*)$. Therefore, we can define the estimated cumulative incidence function for cause k^* at time t^* as follows:

$$\widehat{F}_{k^*}(t^*|\mathbb{X}^*) = \frac{o^*_{k^*,t^*}}{1 - \sum_{k\neq\emptyset}\sum_{n\leq t^*_{j^*}} o^*_{k,n}} \tag{4.1}$$

4.3. Loss function

The optimization process in the Dynamic DeepHit model consists in the minimization of a loss function that regards time dependent variables and right-censoring. It is the sum of three sub

loss functions ($L1 + L2 + L3$). L_1 is the negative log-likelihood of the joint distribution of time-to-event, which enables to handle right-censoring data.

$$L_1 = -\sum_{i=1}^{N} \mathbb{I}(k^{(i)} \neq \emptyset) \cdot log\left(\frac{o^{(i)}_{k^{(i)},t^{(i)}}}{1 - \sum_{k\neq\emptyset}\sum_{n\leq t^{(i)}_{j^{(i)}}} o^{(i)}_{k,n}}\right) + \mathbb{I}(k^{(i)} = \emptyset) \cdot log\left(1 - \sum_{k\neq\emptyset} \widehat{F}_k(t^{(i)}|X^{(i)})\right) \tag{4.2}$$

where:

- $k^{(i)}$ is the competing risk of the i^{th} patient.
- $o^{(i)}_{k,n}$ is the estimated probability given $\mathbb{X}^{(i)}$ of the i^{th} patient having the event k at time $t^{(i)}$

L_2 is the cause-specific ranking loss function :

$$L_2 = \sum_{k=1}^{K} \alpha_k \sum_{i\neq j} A_{kij} \cdot \eta(\widehat{F}_k(s^i + t^i_{J^i}|\mathbb{X}^i), \widehat{F}_k(s^j + t^j_{J^j}|\mathbb{X}^j)) \tag{4.3}$$

where:

- $\eta(a, b) = e^{\frac{a-b}{\sigma}}$.
- $A_{kij} = \mathbb{I}(k^i = k, s^i < s^j)$ is the indicator function that is one if the i^{th} patient has lower probability of CVD event than the j^{th} for the event k.
- $\alpha_k \geq 0$ is a hyper-parameter chosen to trade off ranking losses of the k^{th} competing event, In this work all α_k are chosen equal for the seek of simplicity.

In medicine problems longitudinal measurements of clinical covariates may be highly associated with the occurrence of clinical events. Thus, we introduce an auxiliary task in the shared sub-network. This task makes predictions on the one step-ahead covariates x_{j+1} to regularize the shared sub-network so that the hidden representations preserve information for the step-ahead predictions. L_3 is defined as:

$$L_3 = \beta \cdot \sum_{i=1}^{N} \sum_{j=0}^{J^i-1} \sum_{d\in\mathcal{I}} (1 - m^i_{j+i,d}) \cdot \mathcal{C}(x^i_{j+1,d}, y^i_{j,d}) \tag{4.4}$$

where:

- β is a hyper-parameter.
- $\mathcal{C}(a, b) = |a - b|^2$ for continuous variables.
- $\mathcal{C}(a, b) = -alog(b) - (1 - a)log(1 - b)$ for binary variables.
- $\mathcal{I}$ is the set of time-varying covariates on which we aim to regularize the network.
- $m^i_{j+i,d}$ indicates if the covariate d at time t_{j+1} for the i^{th} patient is missing.

5 | HIV patients dataset

In this chapter we introduce the clinical dataset on which we will apply the methods explained in the previous sections. Data from 4512 patients affected by HIV were collected at IRCCS San Raffaele from 1998 until today. Data from before 1998 are not analysed because the ART was not a combination of the inhibitors yet and the protease inhibitors were not in use. For each patient medical information was registered including demographic variables (e.g., sex, race, age, etc.), clinical parameters (e.g., viremia, cholesterol, etc.) and time-exposure to ART. We defined the period of observation as the one from the beginning of the anti-viral therapy until the occurrence of a cardiovascular event. Specifically we considered only cardiovascular diseases that occurred within 15-years from the beginning of the follow-up. Patients with a CVD event occurred after a follow-up of 15 years were considered as censored data.

5.1. Time to event and censoring

The target variable used in this work is the time interval that goes from the start of the ART at the hospital to the occurrence of a CVD event, if any, or to the censoring moment. For each patient that has experienced an event it, the type of CVD is specified, eight different types are considered:

- Angina
- Cerebral ischemia
- Myocardial infarction
- Ischemic cardiopathy
- Revascularization
- Peripheral ischemia
- Arteriopathy
- Ischemic heart disease

For the seek of simplicity and for the limitation due to sample size, all these different types of diseases have been considered as a unique event.

5.2. Covariates

The IRCCS San Raffaele supplied us the clinical history of each patient from the beginning of the ART to the event/censoring. The clinical history is the collection of all variables measured at

each visit. Thus for every patient we have a certain number of observations. For each observation we have the time that has passed from the beginning of the ART, the time that has passed from the previous visit, whether he/she has had the event in that particular time interval and all the covariates supplied by the hospital. The distribution of the number of visits of each patient is reported in Figure A.20, with an the average of 32 visits up to a maximum of 456.

We have selected 24 variables from all those supplied by the IRCCS San Raffaele. These variables were selected considering the proportion of missing data and consulting clinicians and doctors. Some covariates consist of demographic information about the patient such as sex, race and age. Other variables are clinical measurements, related to CVD, such as the level of cholesterol and the level of creatinine. Furthermore, among these covariates, 5 regard the ART of the patient. In particular, the cumulative year of exposure to drugs each classified by the stage of the cycle they inhibit: the non-nucleotide reverse transcriptase inhibitors (NNRTIs), the nucleoside reverse transcriptase inhibitors (NRTIs), the protease inhibitors (PIs) and the integrase inhibitors (INIs). Moreover we considered as a categorical variable whether the patient has started the ART before or after 2007 since in that year therapies changed. It is noted that these variables are correlated since the ART consists of a combination of a triple-drug therapy.

Among the 24 covariates, 9 variables are binary, 3 categorical and 12 numeric. Moreover 6 are time-independent and the remaining ones depend on time, among the 9 binary variables, 4 are step functions: the presence of a tumor, of diabetes, of AIDs and whether the patient is hypertensive. All descriptions and information are reported in Table 5.1. Basic statistics of categorical variables are reported in Table 5.2 and the numeric ones in Table 5.3.

Name	Description	Type	Time-dependency
Sex	The sex of the patient	Binary	Fix
Race	The race of the patient	Binary	Fix
Year ART	Whether the anti-retro viral therapy have started before 2007	Binary	Fix
Fdr	The factor of risk that could have caused the contraction of HIV	Categorical	Fix
Age	The age of the patient updated at each visit	Continuous	Time-dependent
Diabetes	Whether the patient has been diagnosed with diabetes at each visit	Binary	Step function
Hypertension	Whether the patient has been diagnosed with hypertension at each visit	Binary	Step function
Tumor	Whether the patient has been diagnosed with a tumor at each visit	Binary	Step function
Aids	Whether the patient has been diagnosed with AIDS at each visit	Binary	Step function
CD4	The level of CD4 at each visit	Continuous	Time-dependent
Cholesterol	The level of cholesterol at each visit	Continuous	Time-dependent
Viremia	The level of viremia at each visit	Continuous	Time-dependent
Creatinine	The level of creatinine at each visit	Continuous	Time-dependent
Triglycerides	The level of triglycerides at each visit	Continuous	Time-dependent
AST	ASpartate aminoTransferase	Continuous	Time-dependent
ALT	ALanine Transaminase	Continuous	Time-dependent
Platelets	The number of platelets at each visit	Continuous	Time-dependent
HCV	Whether the patient has been diagnosed with hepatitis C at each visit	Categorical	Fix
HBV	Whether the patient has been diagnosed with hepatitis B at each visit	Categorical	Fix
Hb	The level of hemoglobin at each visit	Continuous	Time-dependent
PI time	The cumulative years of protease inhibitor drug exposure	Continuous	Time-dependent
INI time	The cumulative years of integrase inhibitor drug exposure	Continuous	Time-dependent
NRTI time	The cumulative years of nucleotide reverse transcriptase inhibitor drug exposure	Continuous	Time-dependent
NNRTI time	The cumulative years of non-nucleotide reverse transcriptase inhibitor drug exposure	Continuous	Time-dependent

Table 5.1: Description of the covariates.

Variable	Level	Percentage
Sex	F	0.2
	M	0.8
Factor of risk	MSM	0.43
	Other	0.57
Race	White	0.91
	Other	0.09
HCV	Yes	0.05
	No	0.73
	NaN	0.22
HBV	Yes	0.02
	No	0.33
	NaN	0.65
Year of ART	Before 2007	0.62
	After 2007	0.38
AIDS	Yes	0.17
	No	0.83
Diabetes	Yes	0.07
	No	0.93
Hypertension	Yes	0.31
	No	0.69
Tumor	Yes	0.13
	No	0.87

Table 5.2: Binary and categorical covariates, the last four are time-dependent, the others are time-invariant.

	Age	CD4	Cholesterol	Viremia	Hb	PLT	triglycerides
mean	45.668	610.032	183.240	0.325	14.209	227.848	142.465
std	10.634	340.854	45.121	0.217	2.683	106.393	88.261
min	0.380	0.000	0.200	0.000	4.200	0.000	0.000
25%	38.231	372.000	154.000	0.106	13.200	179.000	84.000
50%	45.152	574.000	181.000	0.409	14.600	221.000	118.000
75%	52.194	799.000	210.000	0.431	15.600	265.000	172.000
max	91.206	4595.000	2357.000	0.905	385.000	800.000	500.000
	Creatinine	ALT	AST	$\text{INI}_t ime$	$\text{PI}_t ime$	$\text{NRTI}_t ime$	$\text{NNRTI}_t ime$
mean	0.916	47.039	36.502	0.472	1.770	3.578	1.373
std	0.385	129.520	78.572	1.465	2.934	4.089	3.005
min	0.000	2.000	2.000	0.000	0.000	0.000	0.000
25%	0.750	20.000	19.000	0.000	0.000	0.032854	0.000000
50%	0.880	28.000	25.000	0.000	0.000	2.184	0.000
75%	1.030	43.000	34.000	0.000	2.565	5.787	0.602
max	10.000	13019.000	8200.000	101.494	22.485	101.494	22.067

Table 5.3: Numerical covariates. These variables are all time-dependent.

6 | Univariate analysis: Kaplan-Meier curves

In this chapter we show the most significant results in terms of the KM estimators, see Section 1.2. The most significant results and the most relevant variables are analysed here, for further details, the rest of the analysis is reported in Appendix A. As already anticipated, the dataset has three different types of variables: categorical fixed-time, binary time-dependent and continuous time-dependent.

6.1. Time-invariant categorical variables

In this section, we show the KM curves relative to the categorical fixed covariates to investigate the relationship between different groups of patients and the probability of CVD event across time. The probability of having a CVD is well explained by the KM estimator curves. In particular, for categorical variables, we can visualize the curve of each group. By looking at the curves we explore of the behaviour of each group while with the help of the log-rank test we can investigate if the curves are significantly different. We report here only the KM curves relative to those covariates that result to be significant. An important variable that can be used in our future models is the beginning year of the ART. It has two levels, before 2007 or after. The two classes have respectively 2782 (62%) and 1730 (38%) patients. The two curves in Figure 6.1 look highly separated, this suggests that the variable is statistically significant, in fact, the log-rank test has a p-value of 0.035. From the figure we see that patients who have started the ART before 2007 are less likely to experience a CVD.

The next categorical time fixed variable is hepatitis C, which has three classes: present, not present or not known. The KM in Figure 6.2 highlights that hepatitis C has an influence on the target variable, indeed the p-value is 0.012. We can deduce that patients affected by HCV are more likely to suffer a CVD.

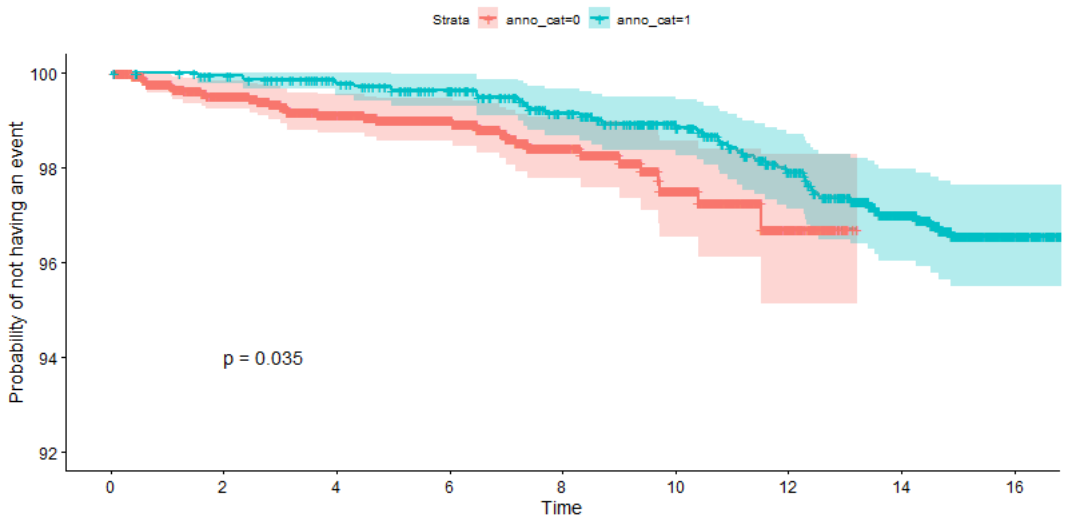

Figure 6.1: KM estimator curves for the year of ART beginning. (1: before 2007; 0: after 2007).

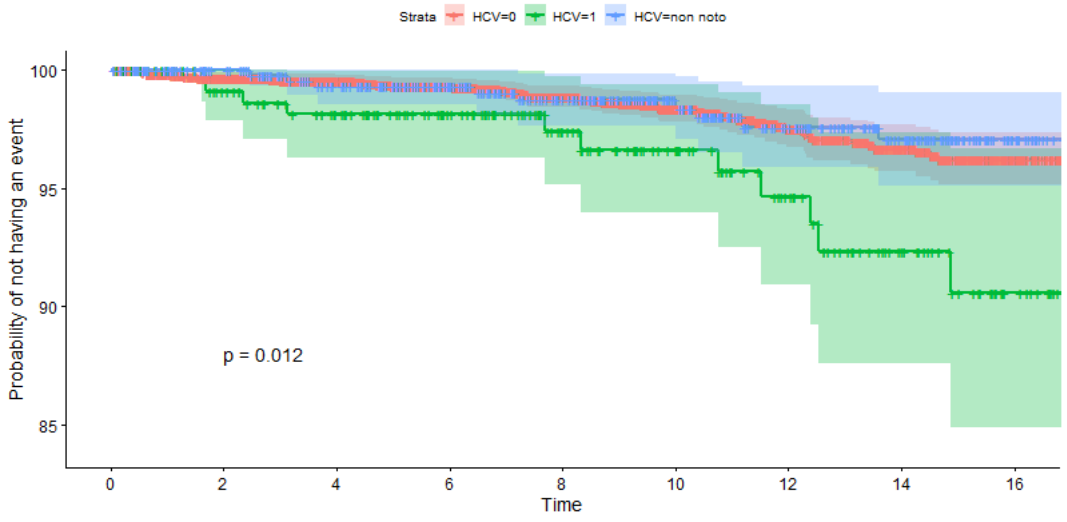

Figure 6.2: KM estimator curves for the hepatitis C. (1: HCV positive; 0: HCV negative; not-known: HCV not known).

Also the race of the patient, that is a binary variable, white or non-white, results to be significant. In Figure 6.3 the two curve appear significantly separated, in fact, the p-value of the log-rank test is equal to 0.09. All other covariate do not result to be associated with different KM curves and are reported in Appendix A.

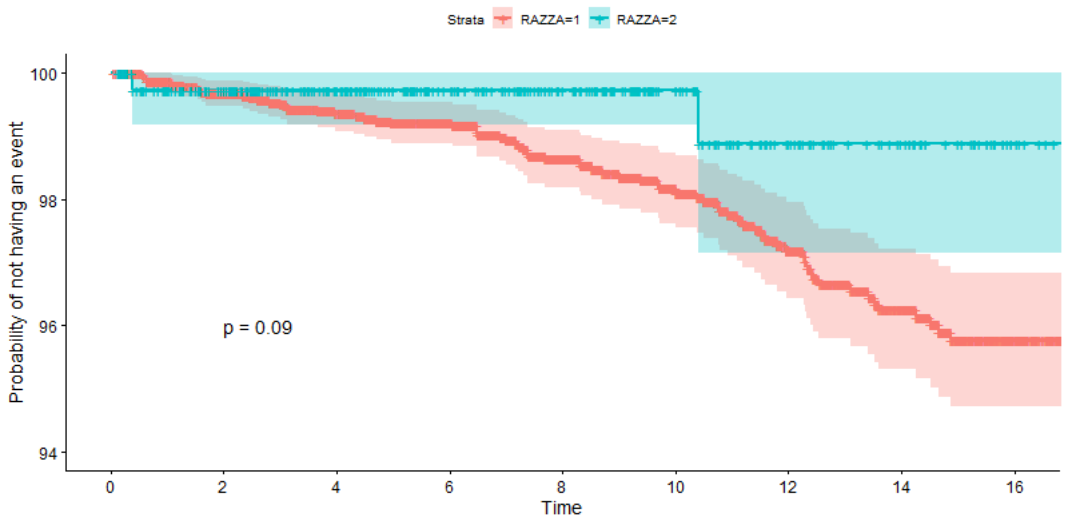

Figure 6.3: KM estimator curves for race. (1: White; 0: Non white).

6.2. Time-dependent variables

In this section we analyse the univariate results of time-dependent variables. There are two types of these variables: binary variables, visualized as step functions, that have zero value until the time of the diagnosis, such as the presence of a tumor or of diabetes, and continuous variables that vary with time, such as the level of viremia. In order to use the KM estimator we have to extend it to the version for time dependent variables. This is a pure descriptive estimate since the log-rank test is not applicable on time-dependent variables.

6.2.1. Binary variables

We have four binary variables: presence of aids, of tumor, of diabetes and whether a patient is hypertensive. All four variables have been considered as step functions, i.e. functions that have 0 value up to the time of the diagnosis and value 1 after.

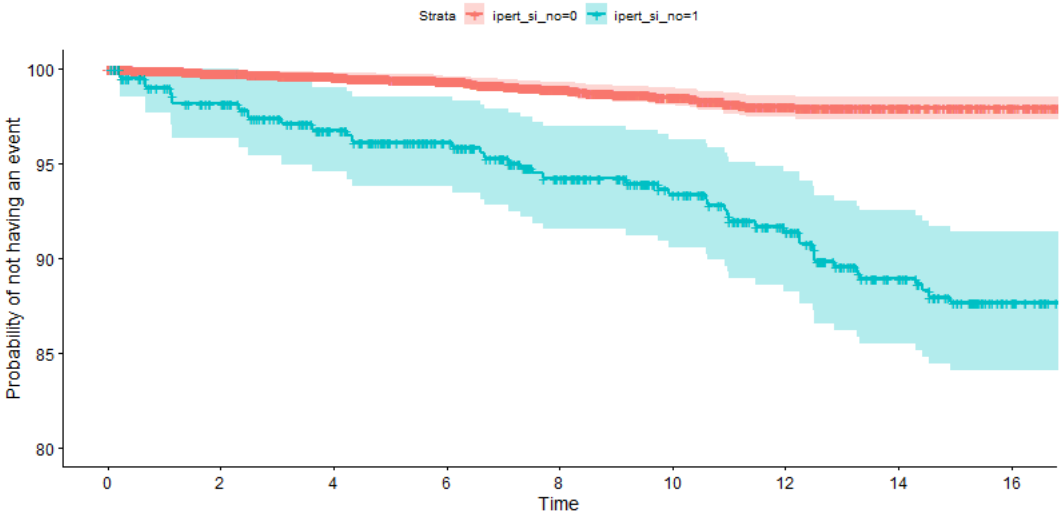

Figure 6.4: KM estimator curves for hypertensive patients.

From Figure 6.4 we can argue that hypertensive patients are more likely to have a cardiovascular disease than non-hypertensive patients. Instead, from Figures A.5 and A.4 in Appendix A, patients with or without a tumor, as well as patients with or without AIDS, do not seem to have different KM estimator curves. Instead, diabetes seems to influence negatively the probability of CVD event, but the low numerosity of patients affected by it makes the confidence intervals very large and consequently, not significant.

6.2.2. Continuous variables

The KM estimator, as we have seen, needs two or more groups to be built, so if we want to analyse continuous covariates we have to discretize them. We have done this using clinical insights and suggestions about critical threshold values for these variables given by the IRCCS San Raffaele. At any time the variable takes the value either 0 or 1 if it is lower or higher of a certain cut-off, respectively.

Here we report only the most significant variables, all the others are reported in Appendix A. The level of CD4, which is a glycoprotein that serves as a co-receptor for the T-cell receptor found on the surface of immune cells, is associated with HIV. CD4 T helpers are white blood cells that are an essential part of the human immune system. If CD4 cells become depleted, like in case of untreated HIV infection, the body is vulnerable to a wide range of infections. There are two different clinical cut-offs below which a patient is considered with low CD4 level which are 350 and 200. Looking at the two KM curves in figures A.7 and A.8in Appendix A we do not have evidence to conclude that CD4 has a correlation with the probability of having a cardiovascular disease. The level of cholesterol is clinically related with cardiovascular diseases

but from the KM estimated curves, using the suggested threshold equal to 250, we can not argue that it is a discriminant variable for our goal (see Figure 6.5 in Appendix A). Even though the number of copies of HIV is clinically related to the target variable, none of the KM estimator curves are related to the probability of CVD, see Figures A.10 with the cut-off at 200 and A.9 with cut-off at 50 in Appendix A. The level of triglycerides, using a cut-off equal to 250, is correlated to the time-to-event. In fact from Figure 6.6 we can conclude that until eight years the two curves are similar while, after, patients with triglycerides higher then 250 are more likely to experiment a cardiovascular disease. Also the level of cholesterol in Figure 6.5 with the cut-off at 250 seems to have two curves significantly separeted.

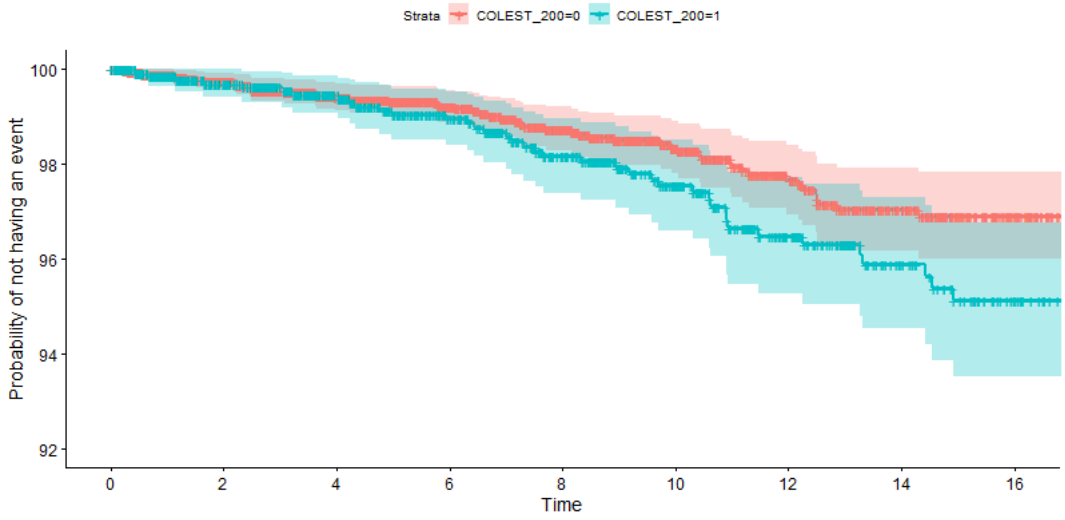

Figure 6.5: KM estimator curves for the level of cholesterol. (1: Cholesterol $\geq$ 200; 0: Cholesterol $<$ 200).

According to the literature [5] [31], the time of exposure to ART inhibitors have a different impact on the CVD risk whether it is a recent exposure, i.e. within 6 months or a later exposure. Therefore we have built KM estimator curves of the exposure time of inhibitors using for the discretizationas cut-off values equal to 6 months. None of the resulting figures suggest a different curve of the probability of CVD event. We report the exposure to PIs in Figure 6.7, the two curves are very similar.

Before presenting the results of the models, it is worth it to underline that the KM estimators of continuous time-dependent variables lose most of the information in the discretization process. In fact, doing this, we are not considering their variability. For this reason some of the last variables may become more important with another model such as the time-dependent Cox or the Dynamic DeepHit.

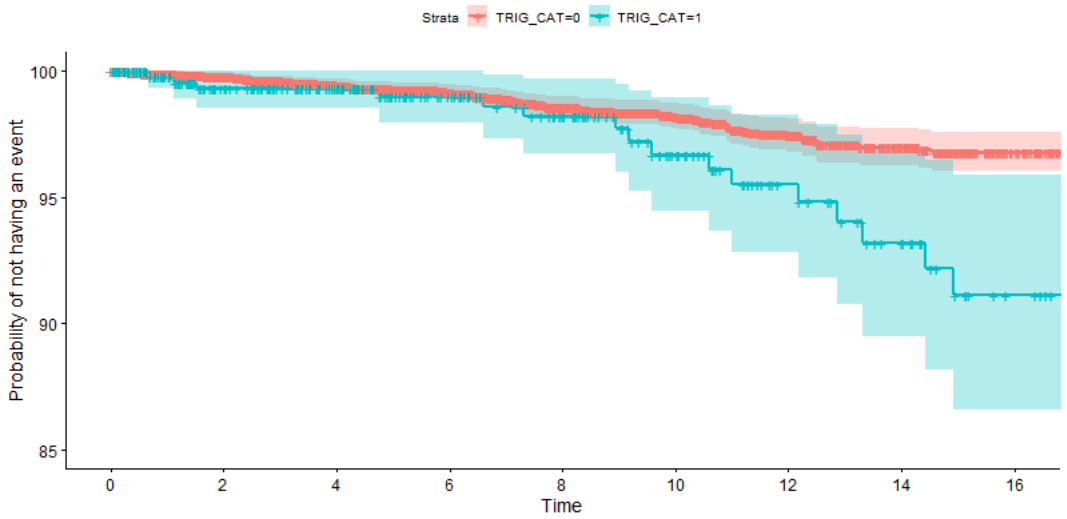

Figure 6.6: KM estimator curves for Triglycerides. (1: Triglycerides ≥ 250; 0: Triglycerides < 250).

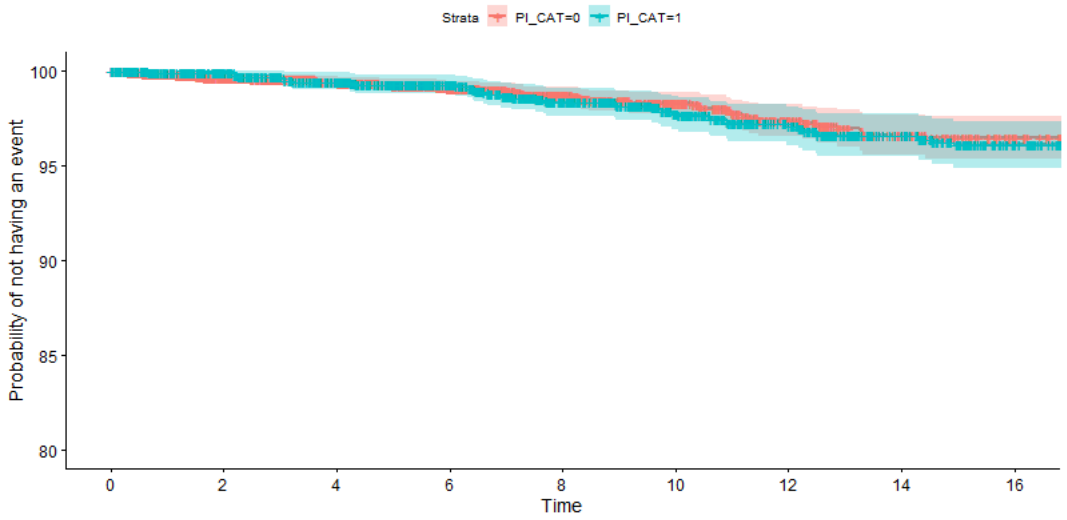

Figure 6.7: KM estimator curves for PIs exposure. (1: Exposure ≥ 6 months; 0: Exposure < 6 months).

7 | Cox and DeepHit models at the baseline: results and comparison

In this chapter, we present the metrics for the evaluation of the compared methods (section 7.1) and the PFI and Shapley values Additive Explanation methods to interpret the DeepHit results (section 7.3). We then apply the Cox PH model and DeepHit at the baseline data (section 7.3).

7.1. Metrics of evaluation

In order to understand the performance of the models we introduce the metrics that will be used in this work to evaluate the models. The mostly used metric in survival analysis is the C-index. As the name suggests, it is based on the comparison between the predicted probability of CVD event and the actual survival time. If we consider two individuals, we can say that we have a good model if it predicts the individual that has the event later more probable to not experiencing the event. To do so the C-index compares all possible couples of patients and for each one counts 1 if it is concordant with the true survival time and 0 if it is discordant. Being concordant means that if individual i is not censored and its time-to-event is smaller than the time-to-event of individual j it predicts that individual i has a lower probability of CVD events. Whether the second patient is censored or not does not influence the result. The C-index is the sum of all comparisons over the total number of possible comparisons. We do not compare two patients if both are censored. C-index is defined as:

$$C - index = \frac{\sum_i \sum_j \mathbb{I}(y_i < y_j)\mathbb{I}(\delta_i = 1)\mathbb{I}(risk_i > risk_j)}{\sum_i \sum_j \mathbb{I}(y_i < y_j)\mathbb{I}(\delta_i = 1)} \quad fori, j = 1, ..., N \tag{7.1}$$

It takes values between zero and one, a zero value represents an inverse predictor, a 0.5 value a random one and the ideal predictor would have a value equal to 1.
The C-index is still the most used metric in survival analysis but it has some drawbacks. Considering for each couple the same value does not give importance to the amplitude of the error, i.e. two patients who have similar hitting time are more difficult to predict as constant than

doing so with two patients with completely different hitting time. Moreover, a model that is not able to predict in concordance two patients far away, in terms of time to event, is not considered a good model but in term of C-index it performs well.
For this reasons some other metrics have been proposed in [4]. Among the others, one consists in using standard regression metrics, in particular the mean squared error on the predicted survival time. The more practical limitation of this new metric is that both the methods used in this work predict only probability of CVD event for every desired time. Neither Cox models nor neural networks based models are able to predict survival time directly. Thus the first task is to extract the survival time from the probability of CVD events. The median time of survival is commonly used, especially in clinical applications, as the predicted time. The median survival time is the moment in which the probability of CVD event reaches 0.5 and actually seems to be a good estimate of the hitting time of the event. The choice of median time is a general guideline but in this application it does not fit well. Indeed, since the proportion of censored patients is 98%, the majority of the KM estimator curves do not reach 0.5. To improve survival time estimation, we use a new threshold that better suites this work, for example the proportion of censored data: 0.98. In order to evaluate this threshold we empirically validated the mean squared error variation with respect to a threshold varying from 0 to 1. In Figure 7.1 the mean squared error variation is shown. Since the lowest mean squared error is actually around 0.98 we decided to use the exact proportion of censored data.

$$\widehat{T}_i = \min_{t \in \mathcal{T}} (\widehat{P}_i(t|X^i) \leq \nu) \tag{7.2}$$

where $\nu = 0.98$ and $\mathcal{T}$ is the prediction time grid.

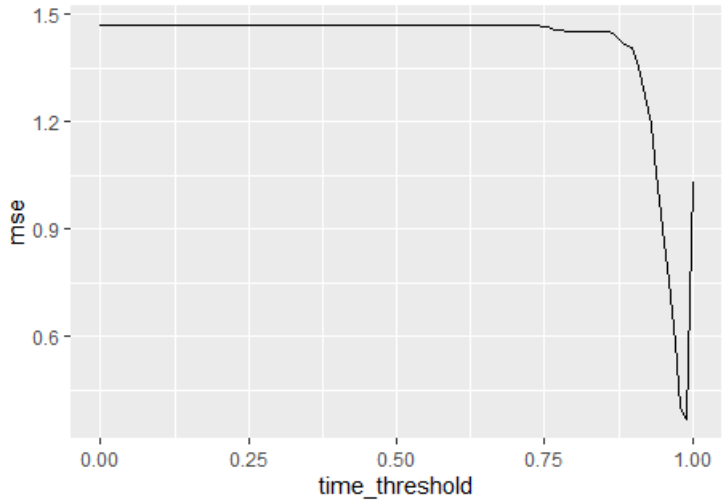

Figure 7.1: Threshold selection for mean squared error.

Since for the two neural network based models it was unfeasible to use a cross-validation approach for computation problems all measurements have been made on a test set of dimension

20% of the total dataset. The two sets, training and test, are built stratified on the number of patients with events, i.e. the proportion of censored patients is the same (98%) in both sets.

The prediction of survival time enables the use of classical regression metrics such as the mean squared error. Therefore we can compare different models performances. The standar mean squared error is defined as:

$$mse(y, \hat{y}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{7.3}$$

where:

- y_i is the i^{th} true target variable;
- $\hat{y}_i$ is the i^{th} predicted target variable;
- N is the number of observations in the test set.

In survival analysis y represents both survival and censoring time. For patients that have had the event the measurement of mean squared error is straightforward, but for censored data y is not the time-to-event but the time of censoring. When we want to evaluate the mean squared error for those patients we have to consider it. Moreover in this work the models were trained on data with time-to-event smaller than 15. For this reason they will not be able to predict a bigger time-to-event and the mean squared error measured over data with higher censoring time is biased.

We suggest a possible solution to these problems through different metrics:

- An estimate of the error in the predicted time-to-event:
  - i) MSE of event individuals: it is the mean squared error measured only on those patients that have had a cardiovascular disease in order to evaluate properly the goodness of fit;
  - ii) MSE of all individuals: it is measured as the mean squared error of all those patients who had the censoring time before 15 years to see how it changes. This tries to evaluate the ability of the model to distinguish censored data from not censored ones. Since the model does not predict further than 15 years we exclude those patients.
- A more suited metric could be defined for this scope. We create a new output prediction that is whether the patient has been predicted to have the event before 15 years or later. Using this binary prediction we can measure:
  - iii) the accuracy;

iv) the sensitivity;

v the specificity.

Note that a majority voting model scores 98% in accuracy and 100% in specificity but 0% in sensitivity (considering the event as positive).

First we will show the results of the two models at the baseline, starting with a full model, that considers all variables. Then, after a features' analysis and selection, we will show the results of the reduced models.

7.2. Permutational feature importance and Shapley values for DeepHit

As already said previously neural networks are black boxes meaning that reading and understanding the structure of the model is extremely difficult. Recently some techniques have been developed to get around this problem with the goal of making this strong models easily readable. The first method we introduce is called *Permutation feature importance*, PFI [21]. This method has the goal of giving a ranking of importance to the features. It works, similarly to the variable importance method of Random Forest [3], as follows:

Algorithm 7.1 Permutation feature importance.

1: Select a metric d
2: Evaluate the dataset $\mathbb{X}$ with d
3: **for** $\forall p \in \{1, ..., P\}$ **do**
4: Permute variable $\mathbf{x}_p$
5: Evaluate the model on the permuted dataset $\mathbb{X}_{perm}$
6: Calculate the importance of variable $\mathbf{x}_p$: $FI_p = \frac{(d-d_p)}{d}$
7: **end for**
8: Return FI

It helps to estimate how much each feature of the dataset contributes to the model's prediction. For this problem we have decided to use the C-index measured on the train set because the predictions on the test set were too unstable. Thus $FI_p = \frac{(c-c_p^{perm})}{c}$.

In Figure 7.2 it is shown the graphic of ordered feature importance estimated with the Permutational algorithm. Although this method gives interesting results that can be used in feature

selection, it does not analyse how the target variable depends on a covariate.
The second method we introduce overcomes this issue. In fact, the *Shapley additive explanation* method [16] gives an idea of how each feature contributes to a particular prediction. To explain the idea behind this tool it is common to compare the problem to a basketball match. Imagine we want to know how much each player contributes to a victory, or even more specifically to the final score. We can evaluate his performance comparing the final score in matches with player A and in the matches without him. In our situation the "final score" is the C-index of the model. We are interested in how each feature affects the prediction of a data point. In a linear model it is easy to calculate the individual effects. A linear prediction for a single observation is:

$$\hat{f}(\mathbb{X}) = \beta_0 + \beta_1 x_1 + ... + \beta_P x_P \tag{7.4}$$

where x_p is the p^{th} feature for that patient and $\hat{f}(\mathbb{X})$ is its prediction. The contribution of the p^{th}, ϕ_p, on the prediction is:

$$\phi_p(\hat{f}(\mathbb{X})) = \beta_p x_p - \mathbb{E}(\beta\mathbb{X}) \quad for p = 1, ..., P \tag{7.5}$$

In neural networks, the contribution of each feature is measured by a game theory based formula that goes beyond our purpose. The Shapley value is a representation of the contribution in the prediction of the prediction of the target variable, therefore a high Shapley value corresponds to an increment in the prediction of the target variable. Through the average of data points Shapley value we have an estimate of the feature importance that can be represented in a similar way to PFI. For each feature we can visualize the Shapley value of every data points, in such a way we understand how different values of the feature, represented by points color from low (blue) to high (red), change the Shapley value. In Figure 7.8 it is reported an example, for each feature it shows the Shapley value of every data points. It is noted that the order of the represented features follows the feature importance that is evaluated as the mean of the Shapley value for each feature.

Moreover we have the possibility to visualize each feature with respect to the target variable, this allow to understand the dependency that exists between the two. In particular we plot every data points, see Figure 7.4, with the Shapley value on the y-axis and the feature on the x-axis. Moreover we may add the information about an other covariate to understand the interaction between the two features, in Figure 7.4 the information about the second feature, is shown with the use of the color of points. Here we can see the relationship between the target variable and the first feature and the interaction between the two features through the colors.

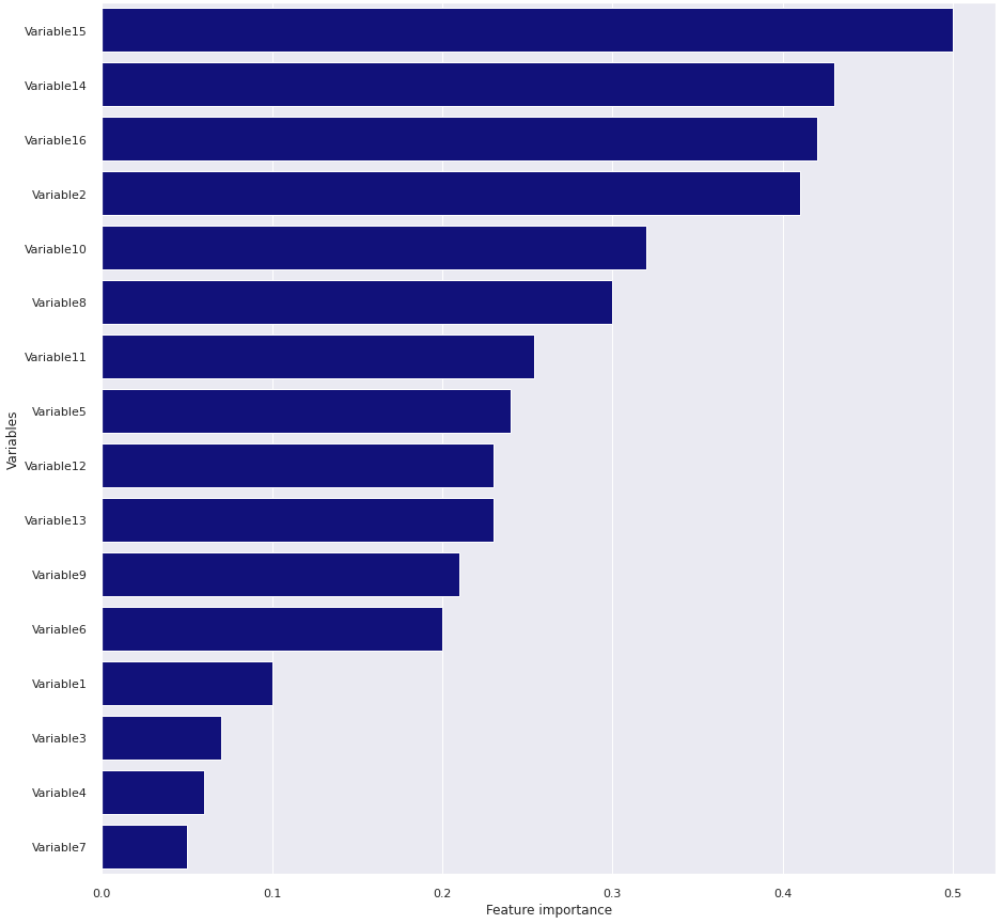

Figure 7.2: Permutation feature importance example.

7.3. Baseline full models

The first two models, Cox model and DeepHit, are built using time-independent data. To build the baseline dataset we have taken all the values of the first visit of each patient, i.e. at the baseline, and we replaced those that were missing with the first available value.

The DeepHit needs a tuning process in order to perform optimally. In this work the architecture of the net used at the baseline has 4 layers in the shared-subnetwork block and 3 in the cause-specific block, each layer has 75 neurons. We have selected the learning rate through multiple runs analysing each time the loss curve and the C-index on the train set. We chose the value of Dropout method as 0.5 and we set the maximum magnitude value for the parameters at 10^{-5}. The weights of each loss functions are 5 for the log-likelihood loss and 1 for the ranking loss.

7.3.1. Interpretation of the models

The Cox non PH model fitted on all the 24 selected variables enables us to understand which are the most significant variables and how they interact with the time-to-event. The coefficients

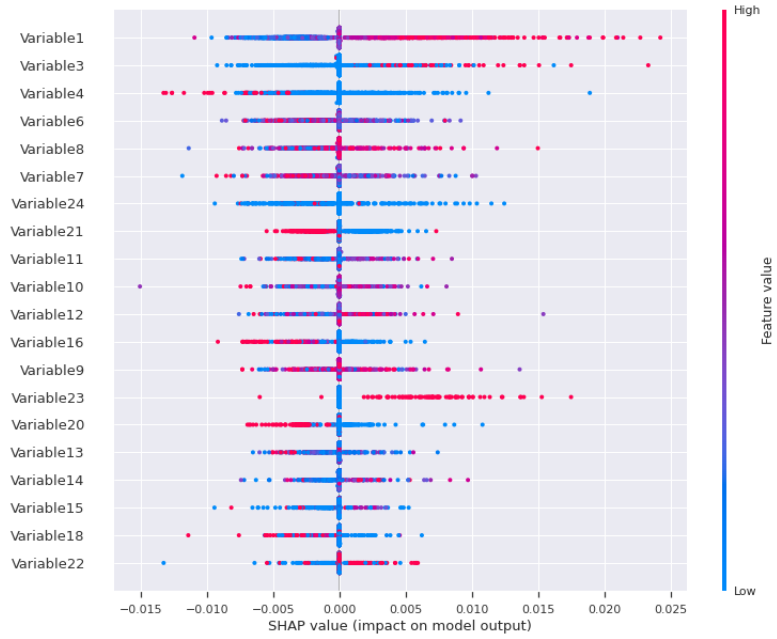

Figure 7.3: Behaviour of variables through Shapley values.

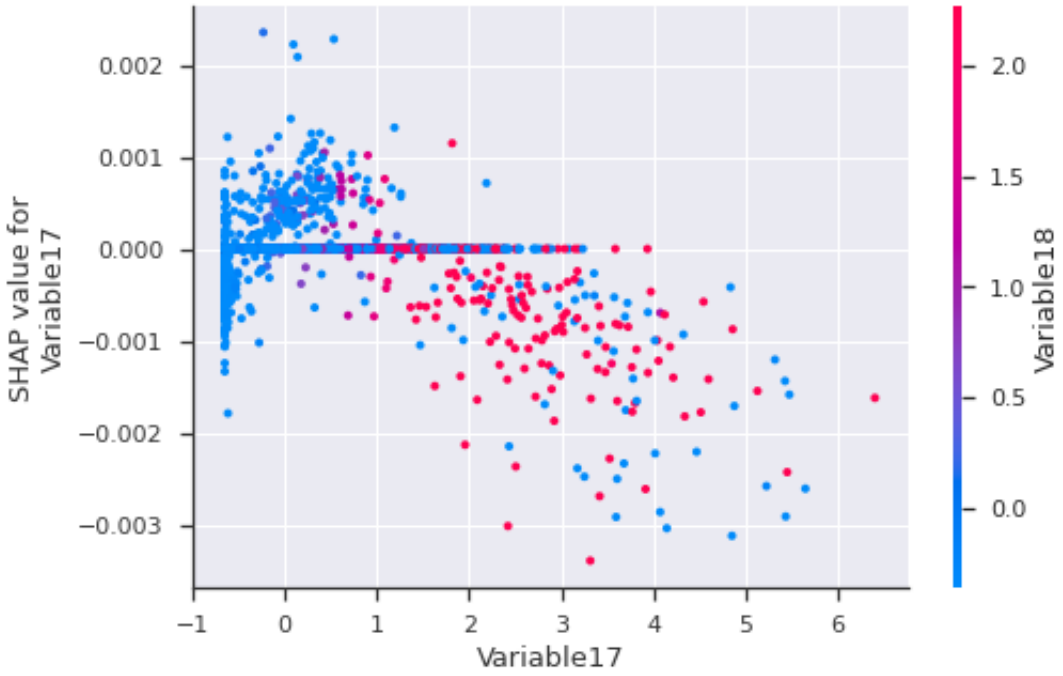

Figure 7.4: Shapley values dependency plot of two variables with interaction.

of the hazard ratio are reported in Figure 7.5 with the confidence intervals and the p-value of the significance of each variable.

The most significant variables are the level of creatinine, whether a patient is hypertensive or not and the age of the patient. From the hazard ratio value we deduce that an hypertensive

patient has a hazard risk of a CVD 2.74 times higher than a non-hypertensive one. A patient one-year-age older than another have 1.04 times higher risk of CVD, while ten-year-older patient have a risk about 1.5 times higher. An increment of 1 in the level of creatinine corresponds with a hazard risk 1.53 times higher. The other variables that are significant for the Cox model at the baseline are the year of beginning of ART, whether the patient have hepatitis C and the level of CD4. Patients that have started the therapy before 2007 have lower risk, 1.75 times, than the others. Patients with hepatitis C have a risk of CVD almost three times higher than those who do not have it. The four variables regarding the ARTs are not significant in the Cox model, but the high correlation between themselves may mask the effects among each others.

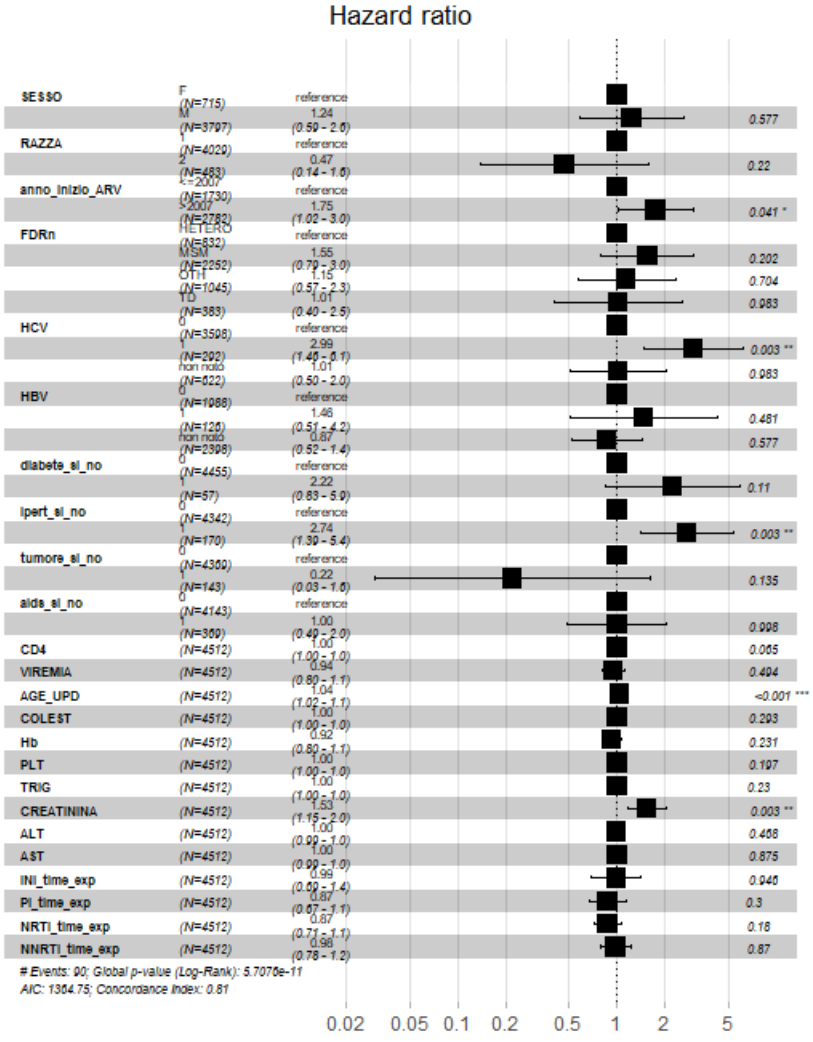

Figure 7.5: Results of the full Cox model at the baseline.

The full Cox model does not satisfy the PH assumption, see Appendix B. Indeed we can see in **??** in Appendix B that the global p-value of the PH assumption test is equal to 0.00028.

Therefore the complete model is not reliable and to get robust conclusions we have to select those variable that are important and that satisfy PH assumption.

The DeepHit model highlights different significant variables. In fact from Figure 7.6, where is reported the feature importance, estimated with the Permutation Feature Importance, of each variable, we can deduce that the most important variables are, in order, the age of the patient, the level of platelets, the presence of HCV and HBV, the year of ART beginning, the cumulative year to exposure of NRTI and the presence of tumor at the baseline.

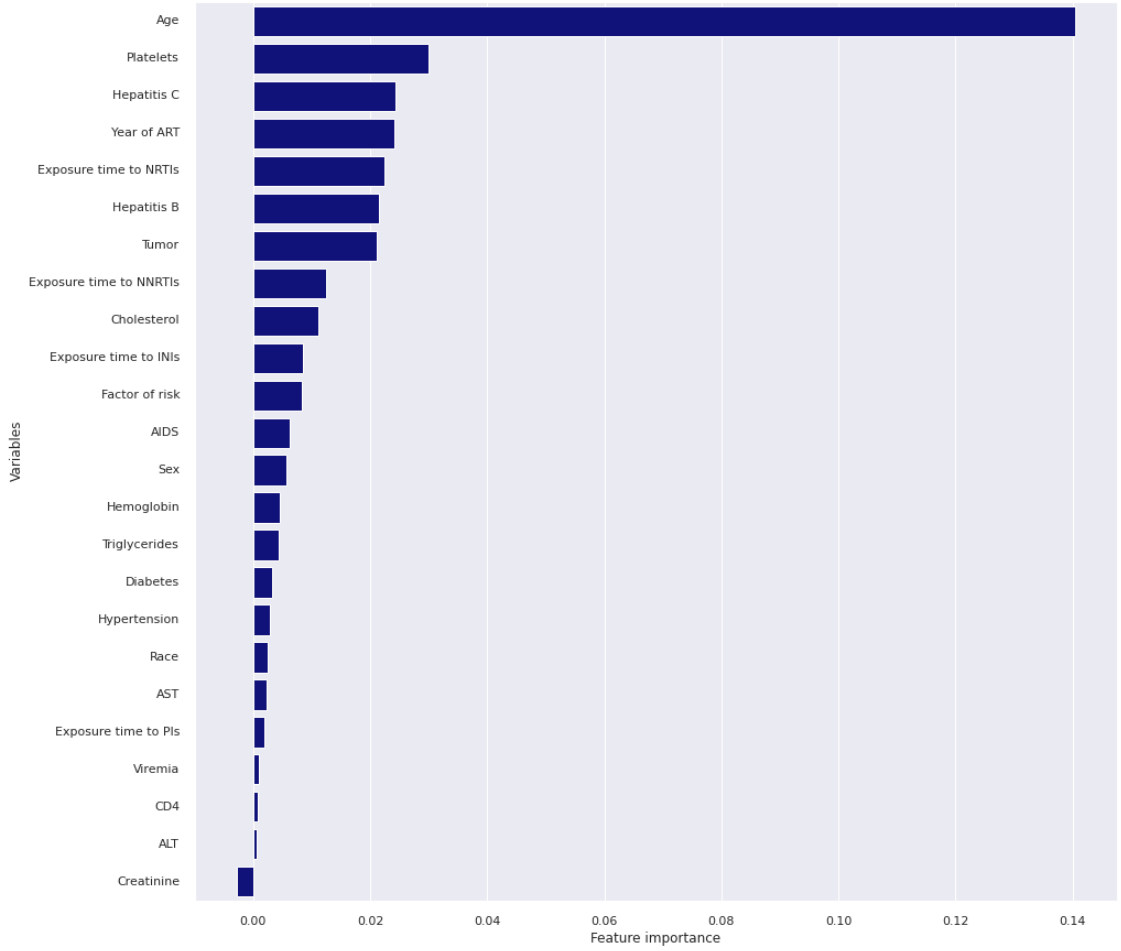

Figure 7.6: Variables importance of DeepHit with PFI of the full DeepHit at the baseline.

The Shapley value estimate of the feature importance highlights some different variables, it is reported in Figure 7.7. In particular, the number of copies of the virus is very important and the NNRTI inhibitors are the most important of all inhibitors. With the Shapley value technique we have a tool that enables us to understand how the covariates interact with the probability of having a CVD, but differently from the hazard ratio of the Cox model, this technique is not quantitative.

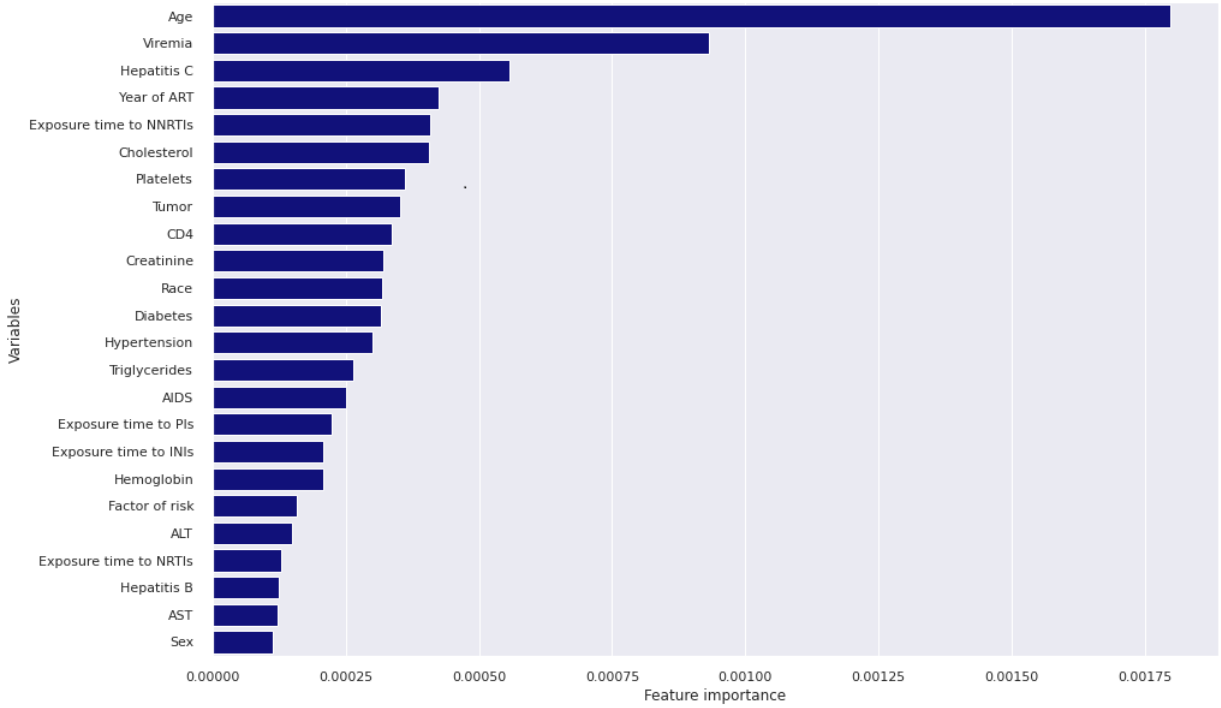

Figure 7.7: Variables importance of DeepHit estimated with Shapley value feature importance of the full DeepHit at the baseline.

In Figure 7.8 it is reported the interaction between the covariates and the time-to-event. The Shapley value, on the x-axis, of each data point represents how a variable, represented through the points color, affects its risk prediction. High Shapley values, right side of the figure, mean high risk of CVD, thus High values of the age, colored in red, contributes to the prediction rising the risk. The result is similar to the Cox model and KM results, in fact, the HCV as well as hypertension are risk factors while the year of ART beginning is a protective one. The ARTs inhibitors variables are all protective, in particular here the NNRTIs are the most important, we recall that the features are order by importance.

7.4. Reduced model

Reducing a model is not an easy task but it can be a strong tool to get more robust performances and also a more readable and simple model. Moreover since the performance of DeepHit is limited by its computational cost, the reduction of the number of variables could result in a big improvement. In addition since in this work only 2% of the observations had experienced a cardiovascular disease it necessary to train a simple model. Also in the evaluation it is important to have enough data in order to make robust evaluations of models.

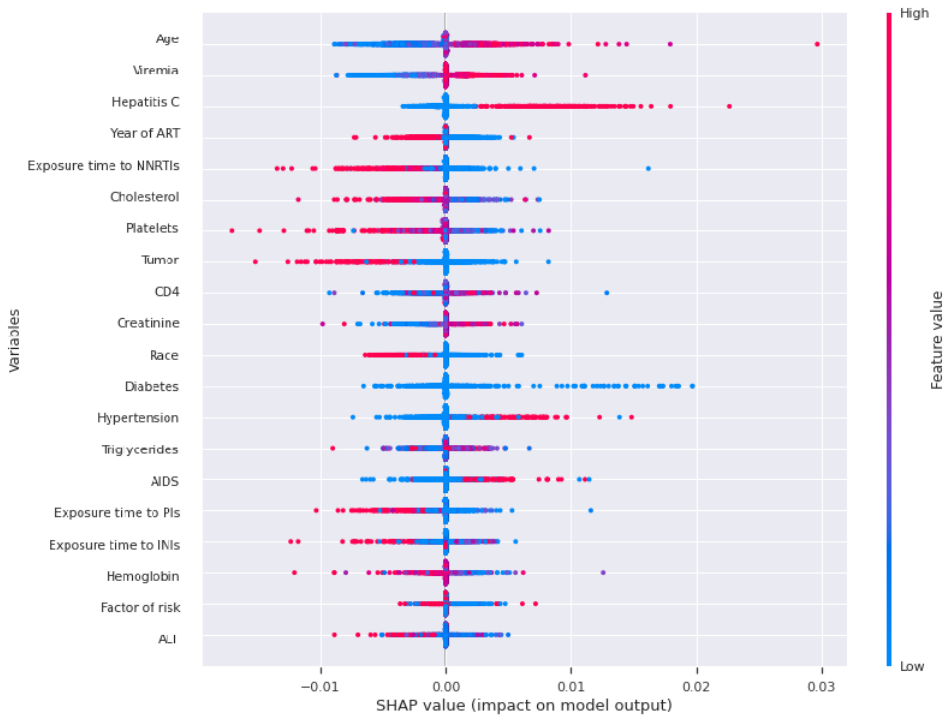

Figure 7.8: Behaviour of variables of DeepHit through Shapley value.

7.4.1. Feature selection

The first models were trained over all 24 variables, a very high number considering that we only have 90 patients uncensored, i.e. experiencing the event. For this reason it is necessary to perform feature selection in order to find a simpler and more performing model, moreover in the case of the Cox model we need to find a model that satisfies PH assumption. In the case of the Cox model we are able to understand the importance of each variable and also how it impacts on first the hitting time using the hazard ratio. Figure 7.5 outlines the most important features and their behaviour with respect to the time-to-event.

With the help of the hazard ratio we built a new Cox model with the most significant variables, which are: the years of exposure to NRTIs, the presence of HCV and diabetes, whether the patient is hypertensive, the age and the level of creatinine. For the Cox model the procedure has been made step by step: we find the less significant variable in terms of p-value, we train a new model without it, and we iterate until the model satisfies PH assumption, all variables are significant and considering the AIC, Akaike Information Criterion of the model. The resulting reduced model has good and robust performances. The inhibitors related variables were forced into the selected variables in order to analyze them. For the reduced Cox model we have selected the following variables, see Figure 7.9.

For the selection of variables for DeepHit we used both the feature importance estimated both with PFI (Figure 7.10) and and with Shapley value (Figure 7.11). Most of the variables are common to the two estimates and both include at list one inhibitors variable: NRTIs or NNRTIs. The process of selection has been made all at once since the computational cost of doing it step by step is prohibitive. The selected variables used in the reduced model are: the year of ART, the presence of hepatitis C, the Age, the number of Platelets and year of exposure to NNRTIs and NRTIs.

7.4.2. Interpretation of the models

The reduced Cox model shows some interesting results. The time of exposure to NRTIs, that was not significant in the full model, is very significant (p-value equal to 0.003) and its hazard ratio indicates that a patient with one year of exposure more than another has 14% risk lower of CVD. It is noted that NRTIs are the only inhibitors in the reduced model beacuse of the correlation between inhibitors.

The other results are similar to results of the complete model, the year of beginning the ART is a protective factor: a patient who has started the ART before 2007 is more than one time and a half less exposed to risk. Note that the reduced Cox model satisfies the PH function (p-value = 0.4), see Appendix B.

Using the Shapley feature importance we have an idea of the ranking of the variable used, see Figure 7.11. For the DeepHit reduced model the most important variable is the age of the patient, followed by the starting year of ART, the viremia, the number of platelets and the exposure time to NNRTIs. From Figure 7.12 we can deduce the relationship between the variables and the time-to-event. As before the age is a risk factor and the starting year of the ART is protective. Here we deduce that the time to exposure to NNRTIs is a protective factor, therefore patients that have been exposed for longer time to this inhibitors class, red points, have a lower risk of CVD with respect to others. It is noted that in this neural network model there are two different classes of inhibitors, in fact, the DeepHit model works also in case of correlated covariates.

A more interesting tool that we can use is the dependency plot between a covariate and the target variable, see section 7.2. In Figure 7.13 we have reported the dependency plot between NRTIs, x-axis, and the Shapley value, y-axis, that represents the risk of CVD. Patients that have a lower exposure time to NRTIs have high Shapley value indicating that they have higher risk. The relationship seems to be almost linear, in particular in the second part of the graph.

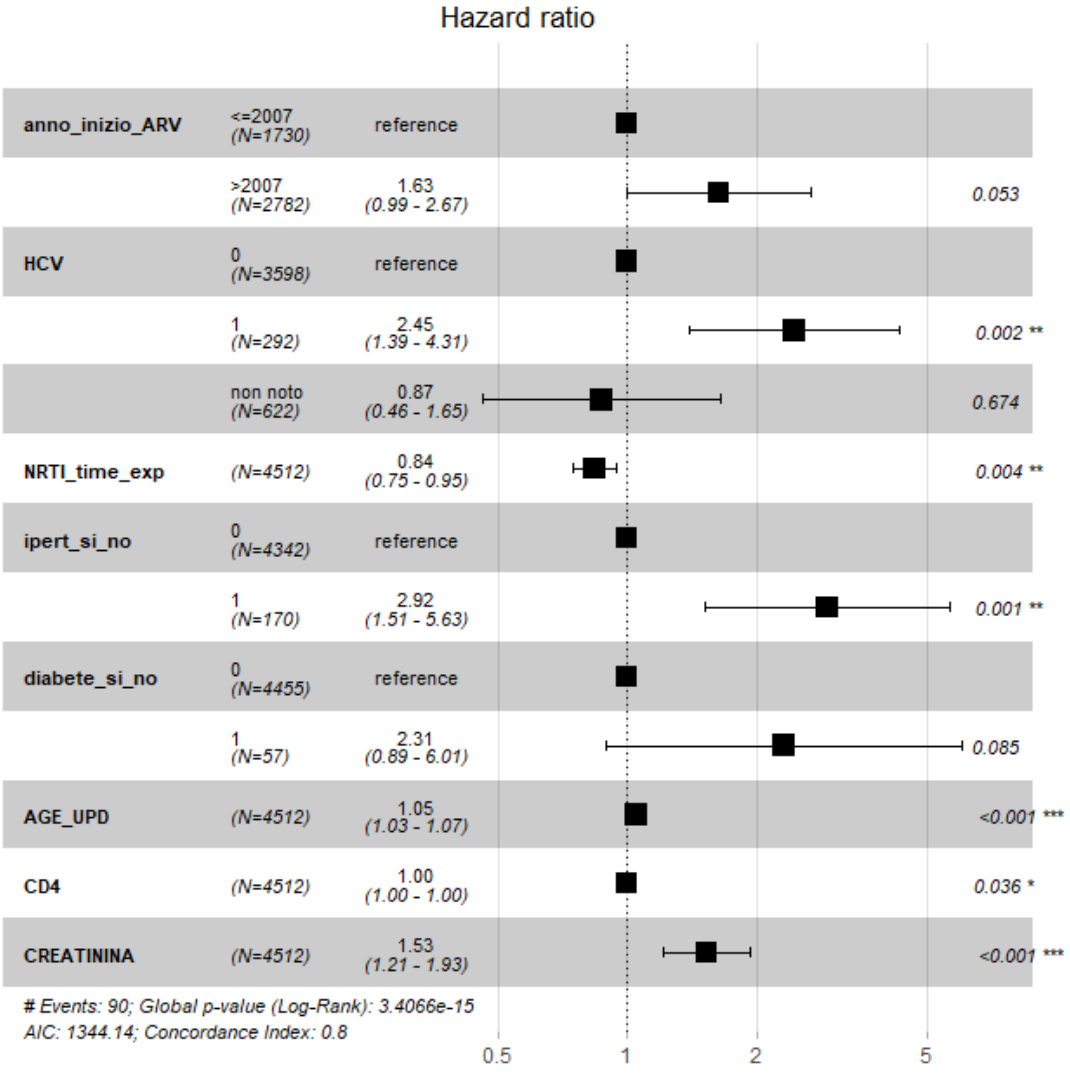

Figure 7.9: Results of the reduced Cox model at the baseline.

The color of the points represent the value of time of PIs exposure, patients that have high value of PIs, in red, usually have also a high value of NRTIs and have lower Shapley values and thus lower risk.

The dependency between NRTIs and Shapley values with the age of the patient as factor is reported in Figure 7.14. In this case we see that old, in red, and young, in blue, patients has a different, but similar, dependency between NRTIs and the time-to-event. In fact from the figure we deduce that old patients with low time of exposure to NRTIs have a very high risk, higher than young in particular, but as the NRTIs exposure rises young patients are more exposed to risk, still being protected, than old patients.

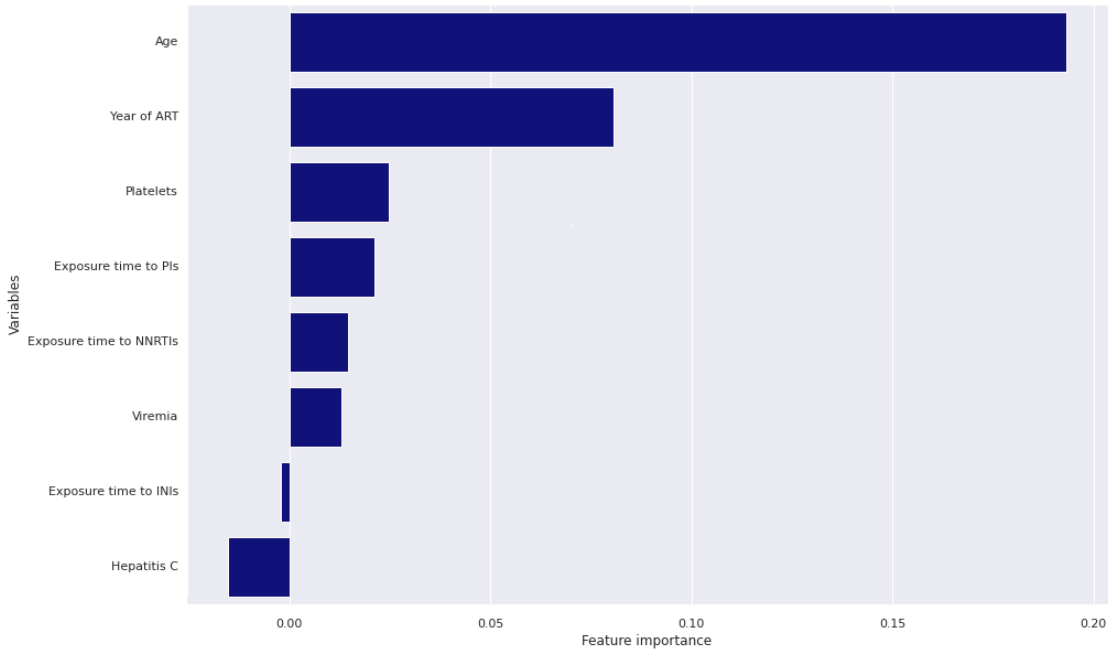

Figure 7.10: Feature importance for the reduced DeepHit model estimated with PFI.

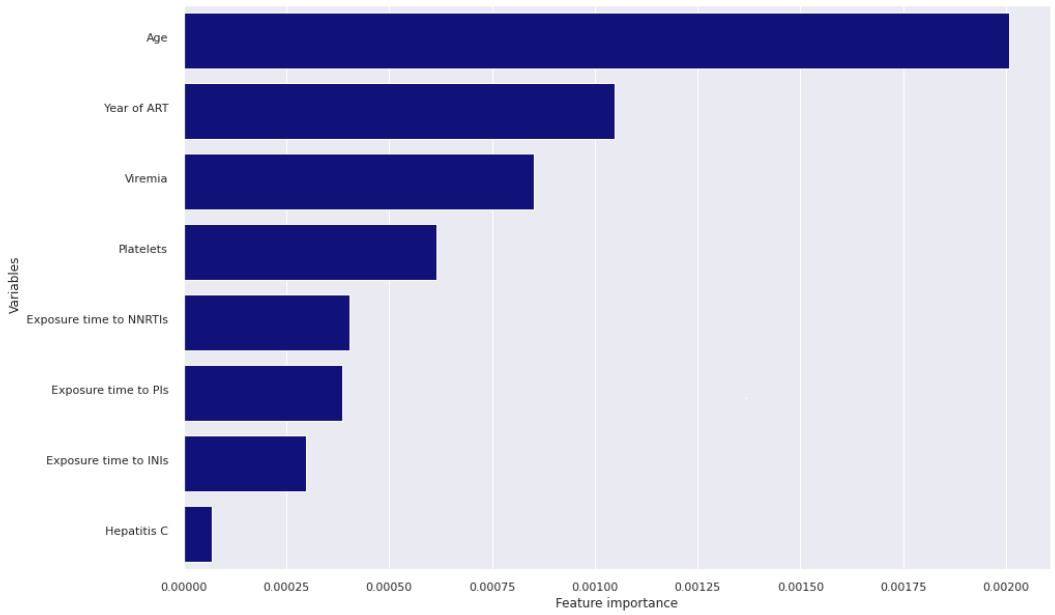

Figure 7.11: Shapley feature importance and dependency for the reduced DeepHit model.

7.5. Evaluation and comparison

The Cox model, fitted over all variables at the baseline, performs well in terms of all the metrics introduced above. In particular, it is the one with highest C-index (0.83) on the training set. On

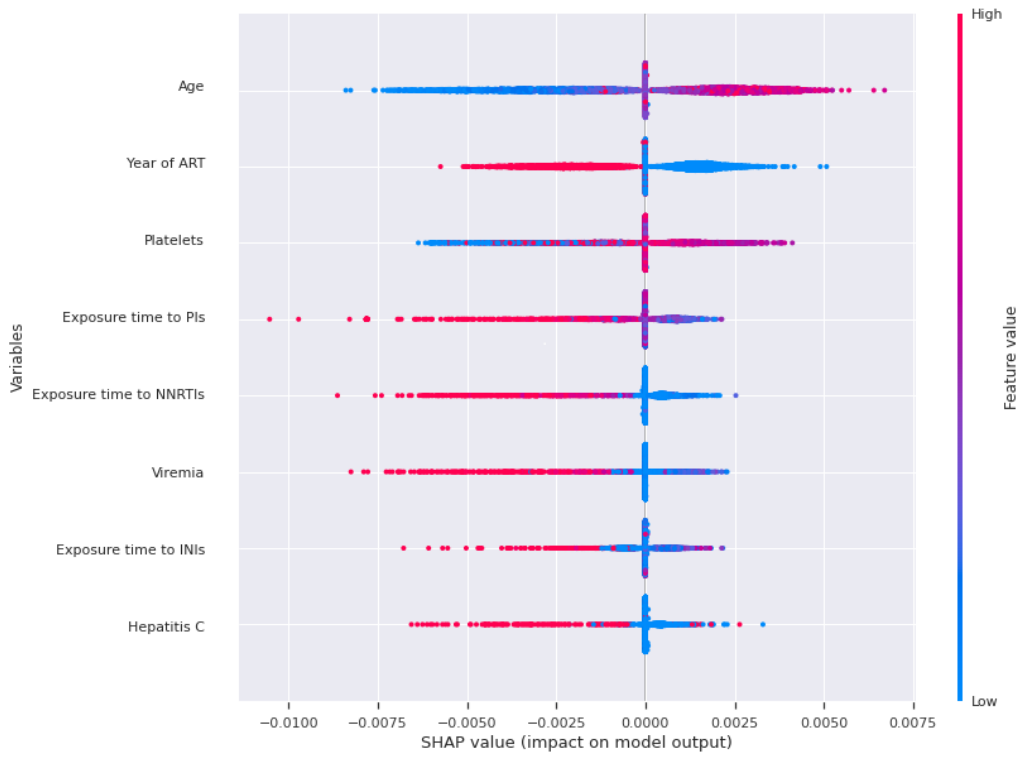

Figure 7.12: The distribution of Shapley values for each feature.

the other side evaluating this model on the test set, which is more important if we want to use the model on a real situation, it decreases more than other models (0.66). In terms of sensitivity (0.94) it is a precise model but the accuracy (0.35) and the specificity (0.34) are small. This means that the model predicts most of the patients to have the event before fifteen years, and thus for the metrics they are counted to have an event.
The two mean squared errors are lower than the ones of the other models, in particular than DeepHit (respectively 26.3 and 41.6 over patients experimenting the event and 39.3 and 47.2 over all patients with event or censoring time before fifteen years). Moreover DeepHit has the lowest accuracy and specificity, in terms of which the Cox model outperforms the neural network approach. What is interesting is that in the training set the machine learning model reaches a low C-index (0.71) but on the test set it is very high (0.74). Note that neural network models were able to gain a C-index value much higher on the training set but have been stopped to avoid bad performances on the test set. All the metrics are reported in **??** and the C-indexes in 7.2.

The reduced Cox model has 8 variables and it is very simple compared to the model with 24. The training C-index remains similar to previous estimates (0.80), even with the loss of information provided by only eight covariates, but the C-index evaluated on the test set rises (0.69) and this

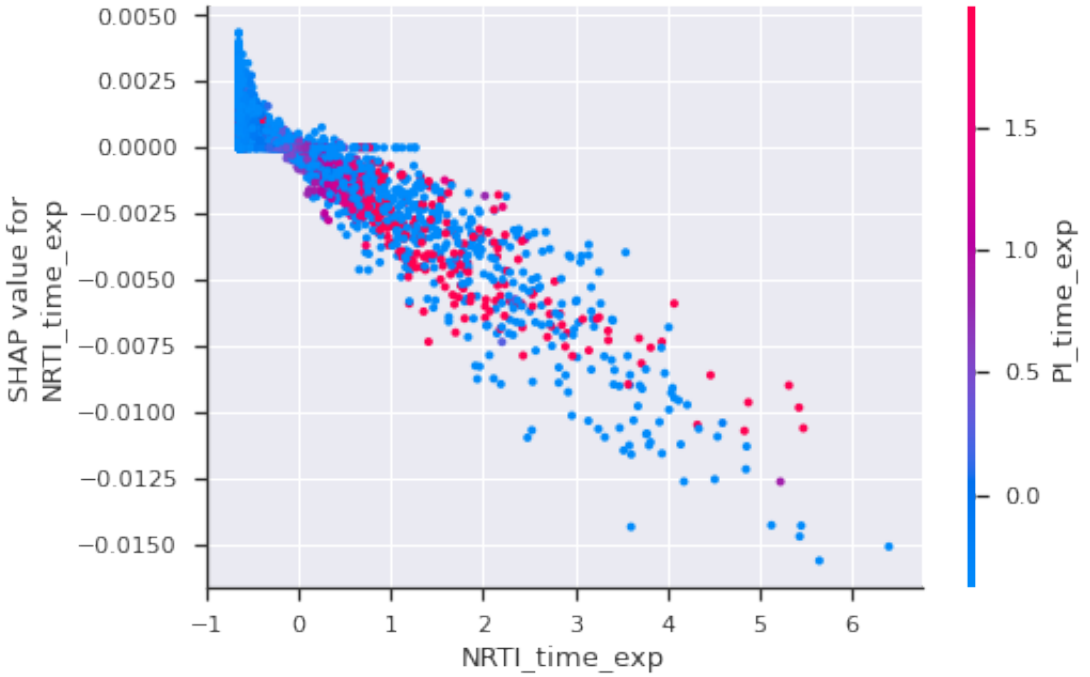

Figure 7.13: Shapley feature importance for the reduced DeepHit model.

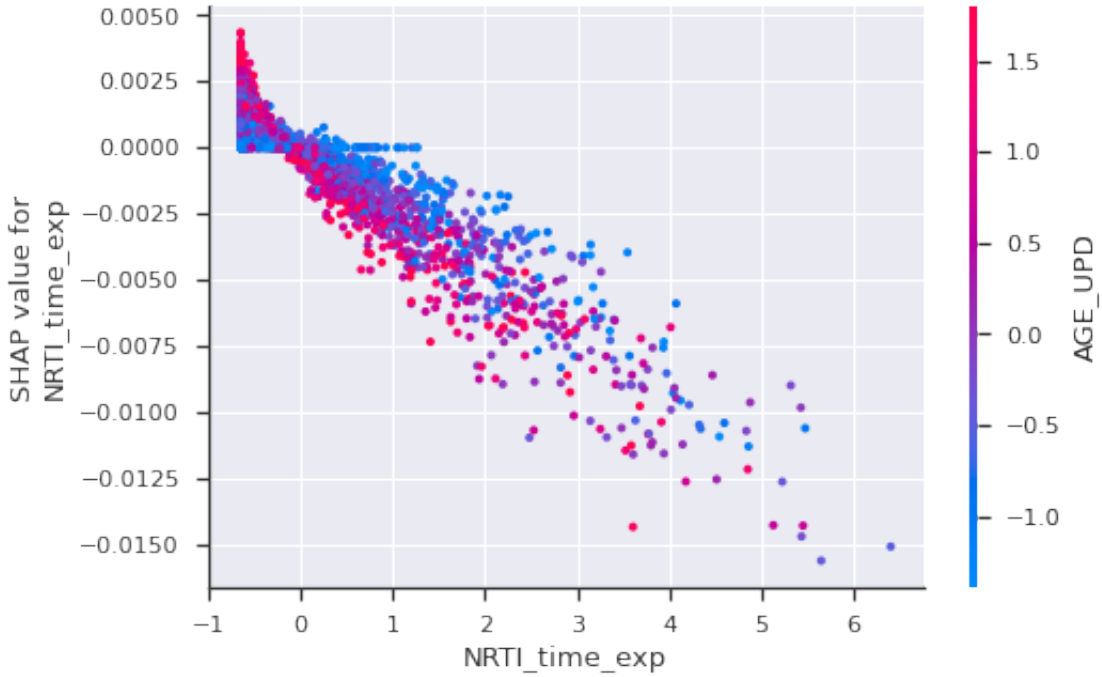

Figure 7.14: Shapley feature importance for the reduced DeepHit model.

indicates that the model is more robust than before towards new data. On the other hand since the time needed to reduce DeepHit step by step was prohibitive it has been reduced at once. The reduced DeepHit does not reach better results in terms of C-index, of which it gets a similar

score in training set (0.80) and a lower score on the test set (0.66). In terms of accuracy (0.442) it rises with respect to the complete model and outperforms also both Cox models (0.280), note that the sensitivity decreases (0.722) while the reduced Cox reaches the maximum (1.00).

Model	Mse event	Mse all	Accuracy	Sensibility	Specificity
Cox	28.286	39.294	0.351	0.944	0.339
Cox reduced	21.285	36.373	0.280	1.000	0.266
DeepHit	41.606	47.231	0.069	0.944	0.051
DeepHit reduced	40.598	57.005	0.442	0.722	0.436

Table 7.1: metrics adopted for the evaluation of the model at baseline.

7.6. Predictive curves

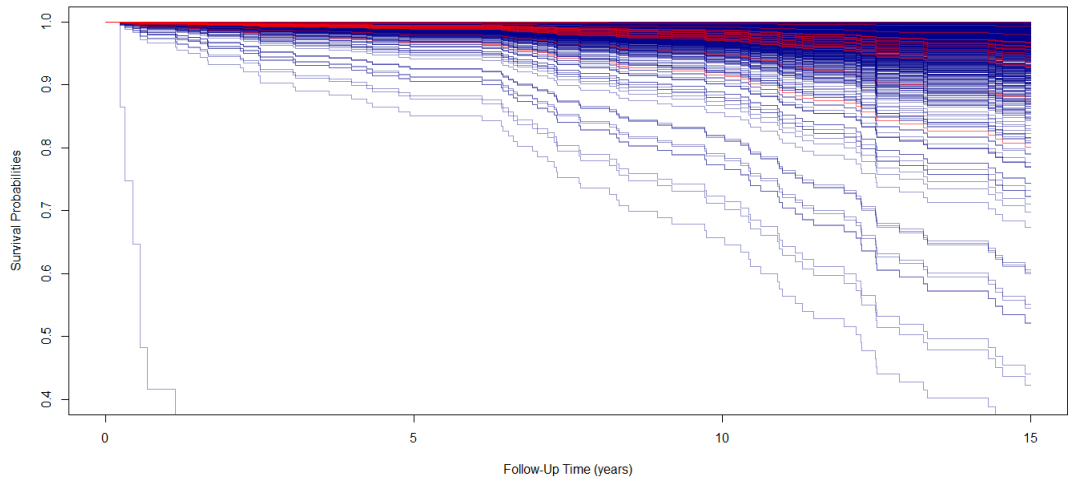

Figure 7.15: Predictive curves of the Cox model at the baseline. Censored data (blue) vs patients with an event (red).

After the analysis of this numerous metrics we want to propose a different way to evaluate our models. The difficulty, that rises when defining a valid metric to evaluate and compare survival analysis models, stands in the the target variable. In fact the predictions of the time-to-event and of the censoring are difficult to be interpreted. Data visualization is a powerful tool that in some situations can tell more than numbers. In this section we propose the comparison between the two full models at the baseline through the representation of curves predicted by the models on the test set. In Figures 7.15 and 7.16 the predictive curves of censored data are represented in blue and the one of patients that had experienced an event in red. A good model should differentiate between the two groups of curves, in particular we expect the red curves to be lower than the blue ones. Such a behaviour indicates that the model recognizes patients at risk and gives them a lower probability of CVD events. Having the curves separated is almost impossible due to the presence of many censored data.

Model	Training C-index	Test C-index
Cox	0.826	0.6603
Cox reduced	0.794	0.6912
DeepHit	0.7113	0.7394
DeepHit reduced	0.8001	0.6564

Table 7.2: C-index for models at the baseline

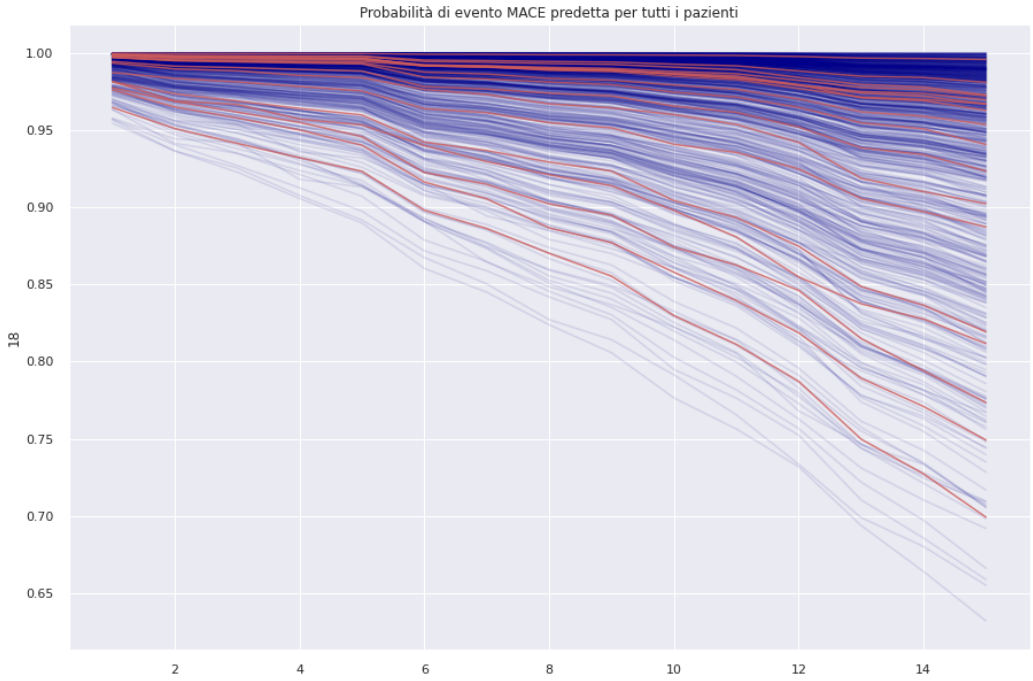

Figure 7.16: Predictive curves of the DeepHit model at the baseline. Censored data (blue) vs patients with an event (red).

The two figures highlight that there is not a different behaviour in the prediction. Even though the models are not able to exactly differentiate between the two type of curves, we can clearly see that in the top there are much more blue curves than reds. Moreover the DeepHit model performs better, in fact in the top of the images there are mostly censored curve and in the bottom mainly red ones.

The curves of the reduced model are reported in the Appendix B and they do not report significant differences.

8 | Time-dependent Cox and Dynamic DeepHit with time-dependent covariates: results and comparison

In this section we will consider longitudinal data models, which are much more complex because they have to handle a bigger amount of data. At the same time they are able to capture relationships that are significant only over time. The introduction of time-dependency is really important in this work because the relationship between the probability of survival and the time-dependent covariates could tell us whether a rapid variation of the number of copies of HIV leads to a higher risk or not. For example we can explore, if, after a certain period of exposure to ART drugs, the risk of cardiovascular disease event rises. The dataset is composed by the same patients of the dataset at the baseline, but for each patient we have the collection of information measured at each visit, reported as multiple observation. For each visit the dataset reports the time from the beginning of the ART and the time from the previous visit, whether the patient has had the event or not and, if so, when it has happened. Moreover all the 24 covariates are reported, measured at the time of the visit. In case of missing value, it may happen that in some visits not all information are registered, we have imputed the most recent value available.

8.1. Full models

We introduce the models with all the 24 covariates and after performing feature selection, we consider simpler models. Here we recall that we have both longitudinal and fixed variables, e.g. the sex of a patient does not change during the follow up and therefore is considered as fixed in time. Most of the covariates are time-dependent variables, some of them such as the number of copies of the virus are continuous, others, such as the presence of aids or tumor, are binary but considered longitudinal being visualized as simple step functions.

The Dynamic DeepHit model selected in this work has 75 neurons for each layer, 3 recurrent

neural network layers, 3 feed-forward layers in the shared block and 3 in the cause-specific block. Higher values made the network too big and, having a low number of observation, it became impossible to make the network learn. The best learning rate, found after many attempts, is equal to $7 \cdot 10^{-4}$, the regularization magnitude is set to 10^{-5} that is the suggested value in literature. The weights of each loss functions selected are 5 for the log-likelihood loss, 1 for the ranking loss and 3 for the prediction loss, we have started trying the suggested values and tried also different ones which resulted to better suite to our problem.

8.1.1. Interpretation of the models

We explain the results of the time-dependent Cox through the hazard ratios and p-values, reported in Figure 8.1. Similarly from before we find that the most significant variables are the starting year of ART, the age of the patient, whether he/she is hypertensive, the presence of HCV. The level of cholesterol and of triglycerides are significant, we recall that these were not significant in the Cox baseline model. The hazard ratio of the level of cholesterol (1.0022) indicates that it is a risk factor. Also the triglycerides are risk factor (with an hazard ratio of 1.0042).

The Dynamic DeepHit can be interpreted through the PFI in Figure 8.2. The computational cost of the Shapley value, where we measure the contribute of each feature on every individuals, is prohibitive with longitudinal data. Therefore we are not able to inspect the time-dependency of the structure of the relationship between the covariates and the time-to-event. In Figure 8.2 are reported the most important variables estimated with the PFI. The results are different from the ones of the time-dependent Cox model, in particular the most important variables are the time of exposure to NNRTIs and PIs drugs. This is possibly due to its ability of capturing non-linear relationships. The other important variables include the time of exposure to INIs inhibitor drugs and the Hepatits C.

8.2. Reduced model

In this section we analyse the two models on a dataset with a reduced set of covariates in order to have clean, simpler and robust models. We perform feature selection using the hazard ration and their p-values for the time-dependent Cox model by a one-step procedure and the feature importance, estimated by the PFI, for the Dynamic DeepHit.

8.2.1. Feature selection

The hazard ratios in Figure 8.1 shows the most significant variables for the Time-Dependent full Cox model. For this model some of the most important features are the same of the model at the baseline. The age of the patient or whether he/she is hypertensive that seem more significant

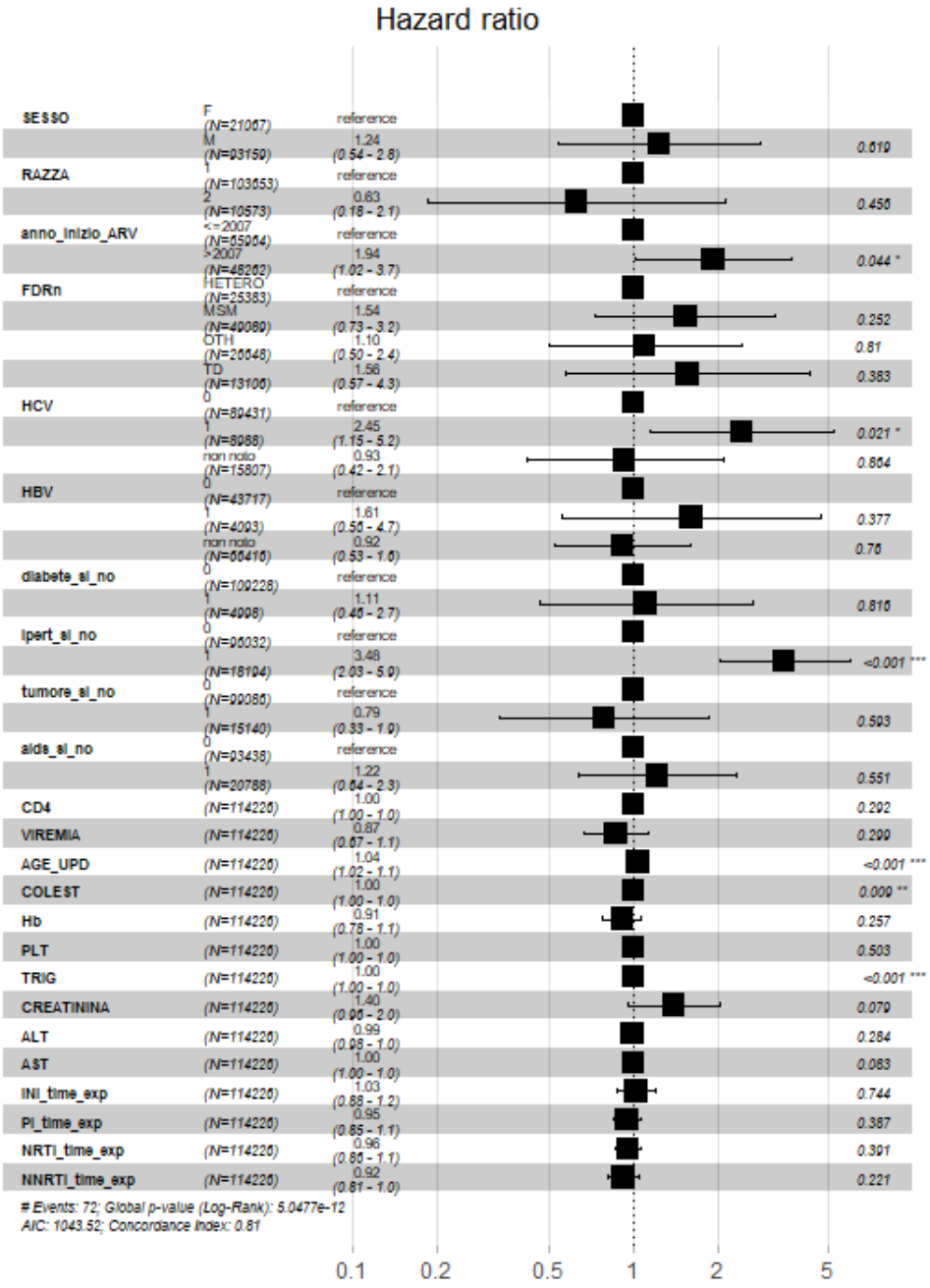

Figure 8.1: Results of the time-dependent full Cox model.

with respect to the models at the baseline, while the number of triglycerides is now one of the most important and was not before. The process of selection of the variables has been made step by step as before, while for the Dynamic DeepHit we have performed it all at once considering the results reported in **??**. As in the case at the baseline, to make feature selection of the DeepHit model we used PFI selection using as metric the C-index measured on the training set. The variables regarding the ART are the most important and since the Dynamic DeepHit model

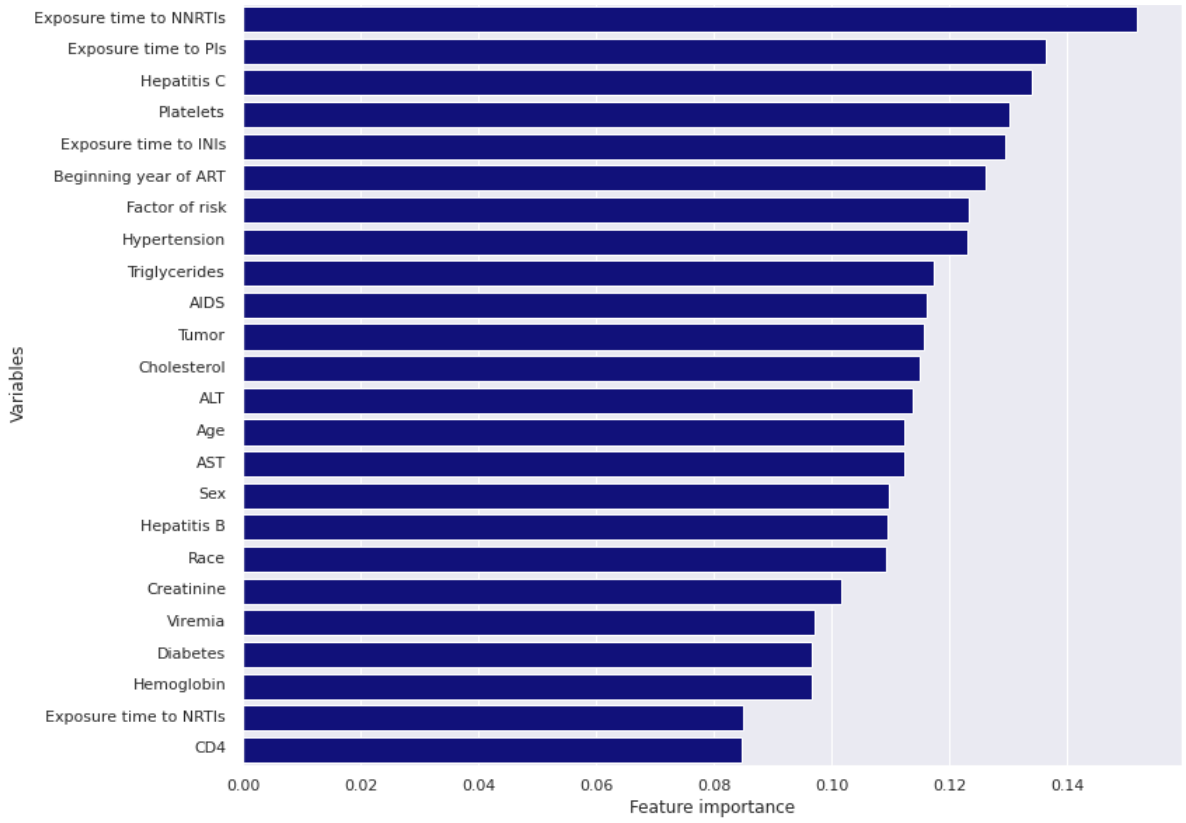

Figure 8.2: Variables importance of Dynamic full DeepHit.

can handle covariates correlated between themselves we have included them all in the reduced model.

The reduced time-dependent Cox model was trained on the following variables: the year of the beginning of the ART, whether the patient suffers of hepatitis C, whether he/she is hypertensive, his/her age, the level of cholesterol and creatinine, the level of triglycerides and the time of exposure to NNRTIs. See Figure 8.3. The Dynamic DeepHit reduced model is fitted on the following variables: the time of exposure to PIs, NNRTIs, NRTIs and INIs drugs, the age, the hepatitis C, the hypertension, the year of the beginning of the ART and the platelets. See **??**.

8.2.2. Interpretation of the models

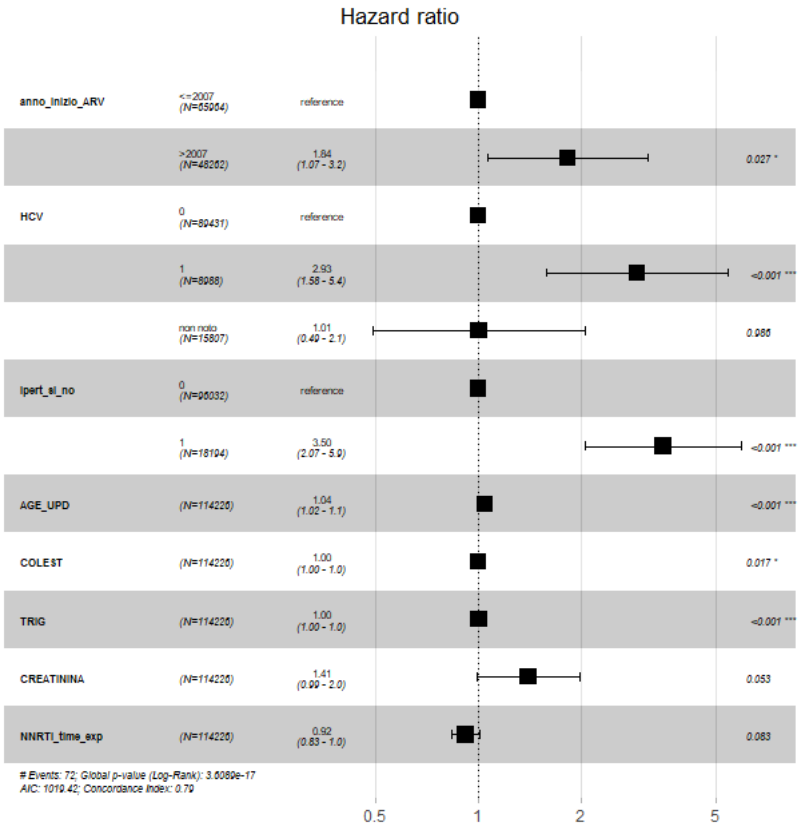

Figure 8.3: Results of the reduced time-dependent Cox model.

The hazard ratios and their p-values are reported in Figure 8.3. From it we get to similar conclusions of the baseline models. In fact, the age, as well as the hepatitis C, the level of the triglycerides and the hypertension (all with p-value < 0.001, is a risk factor and has almost the same hazard ratio of the baseline model (1.042 and 1.0044 respectively). Also the level of creatinine is significative (p-value $= 0.053$) and it a risk factor. As for the baseline models patients who have started the ART after 2007 are almost 2 times more at risk than the others (p-value $= 0.027$). The time of exposure to NNRTIs is a protective factor, in particular, a year of exposure to drugs of this class of inhibitors corresponds to a decrease of 8% in the risk of a CVD event (p-value $= 0.083$).

The feature importance of the reduce Dynamic DeepHit model is reported in Figure 8.4. The result is similar to the complete model: the most important variables regard the ARTs inhibitors, in particular the NNRTIs and the PIs seem to be the most important followed by the hepatitis C and the platelets. These results underline the importance in a predictive model of the information regarding the ARTs, in particular all inhibitors are important in the predictions and also the year of the beginning of the ART results as one of the most important.

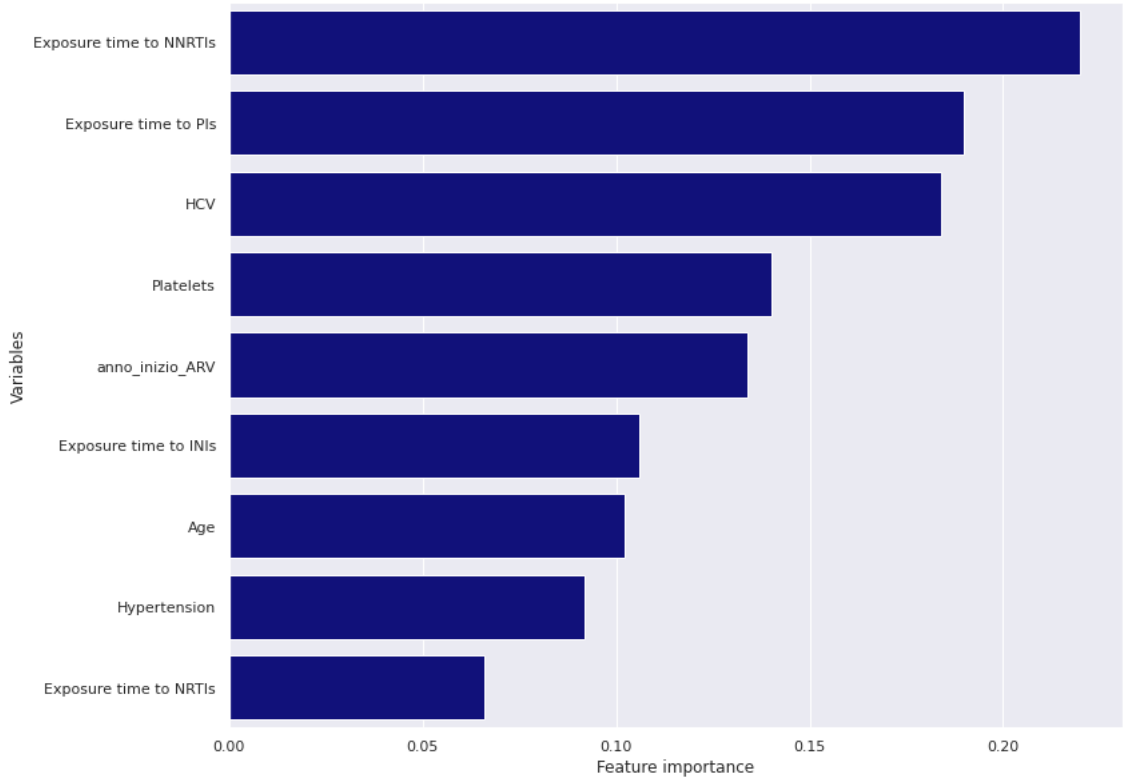

Figure 8.4: Feature importance for the reduced Dynamic DeepHit model estimated with the PFI.

8.3. Evaluation and comparison

All the evaluation metrics are reported in the Tables 8.2 and 8.1. The time-dependent Cox model does not seem to outperform the baseline one in terms of mean squared error, being both, the one measured over uncensored patients (54.411) and the one over all patients (50.510), greater than the model at the baseline. But in terms of C-index it gains better performances reaching high values both on the training set (0.81) and on the test set (0.68), moreover the accuracy (0.740) and specificity (0.749) that it reaches are really high. The only drawback is the sensitivity (0.278), which is the most important for us. It suggests that the model predicts the majority of the patients living more than 15 years, and thus we consider them as censored.

The Dynamic DeepHit needs a note before the beginning of its analysis. Since its training has been really difficult, time consuming and computationally intense it may have not reached its full potential. In fact the training, as well as the evaluation of feature importance and the evaluation of the metrics, has been really long with an average of 16 hours. Considering that it needs more than one or two runs in order to tune the hyper-parameters and that the optimization process works well with multiple random starts, the training becomes a very long process. Even with these difficulties it has reached some interesting results, indeed the C-index evaluated over the

training set and over the test set are similar (repsectively 0.7701 and 0.7713) indicating that this model is robust towards new data. Note that on the test set it has the highest value. All other metrics are lower than those of the time-dependent Cox model and of the baseline DeepHit.

The time-dependent Cox with the reduced covariates, which have been selected step by step, has the same training C-index of the full model (0.80) but a higher value on the test set (0.70). This suggests that the model is more robust towards new data. In terms of mean squared error there are no significant change as well as for accuracy and sensitivity.
The Dynamic DeepHit reduced model does not perform as wellas the complete model. In fact the C-index decreases, with respect to the full Dynamic DeepHit, both on the training set (0.76) and on the test set (0.72). This suggests that all time-dependent covariates add some important information, this reflects the PFI, in Figure 8.2, where all the variables have a medium-high value of importance. All the metrics are reported in **??** and the C-indexes in 8.2.

Model	Mse event	Mse all	Accuracy	Sensibility	Specificity
time-dependent Cox	54.411	50.510	0.740	0.278	0.749
time-dependent Cox reduced	59.272	50.424	0.742	0.333	0.750
Dynamic DeepHit	47.894	37.587	0.147	0.556	0.139
Dynamic DeepHit reduced	49.950	44.367	0.019	0.944	0.001

Table 8.1: Metrics adopted for the evaluation of the time-dependent Cox and dynamic DeepHit models.

Model	Training C-index	Test C-index
time-dependent Cox	0.809	0.6782
time-dependent Cox reduced	0.7998	0.7002
Dynamic DeepHit	0.7701	0.7713
Dynamic DeepHit reduced	0.7601	0.7228

Table 8.2: C-index for time-dependent Cox and Dynamic DeepHit models.

8.4. Predictive curves

Looking a the predictive curves of the models with time-dependent, these models do not seem to predict better than before the probability of not having an event. In **??**, the predictive curves of the time-dependent Cox model are represented, in red the patients that had an event and in blue the ones censored. Here the two groups look similar, we cannot differentiate between them but the trend of the curves of patients that had the event (red) is clearly lower than the censored data curves (blue).

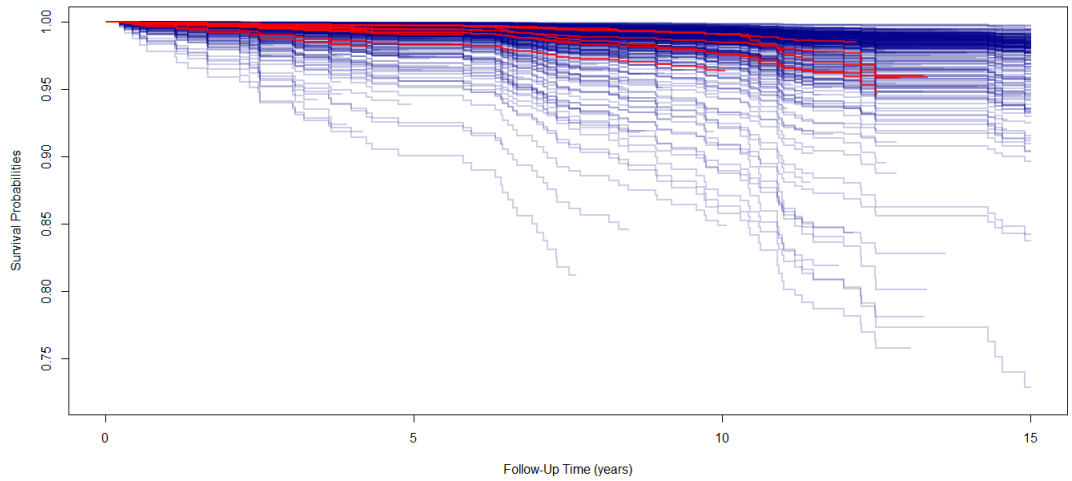

Figure 8.5: Predictive curves of the time-dependent Cox model. Censored data (blue) vs patients with an event (red).

The Dynamic DeepHit curves have a strange behaviour, see **??**, we are not able to distinctly recognize a difference between red curves and blue ones. This behaviour may be affected by the fact that after the censoring moment a patient does not have any further information, so the model have the only information of how much time has passed from the beginning of the ART. Time is a risk factor, i.e. as time passes the probability of having an event increases, and without any other information the model predicts the patient with very low probability. This behaviour is not visible in the curves of the time-dependent Cox model because it do not try to predict after the censoring moment, i.e. the curves stop at certain points.

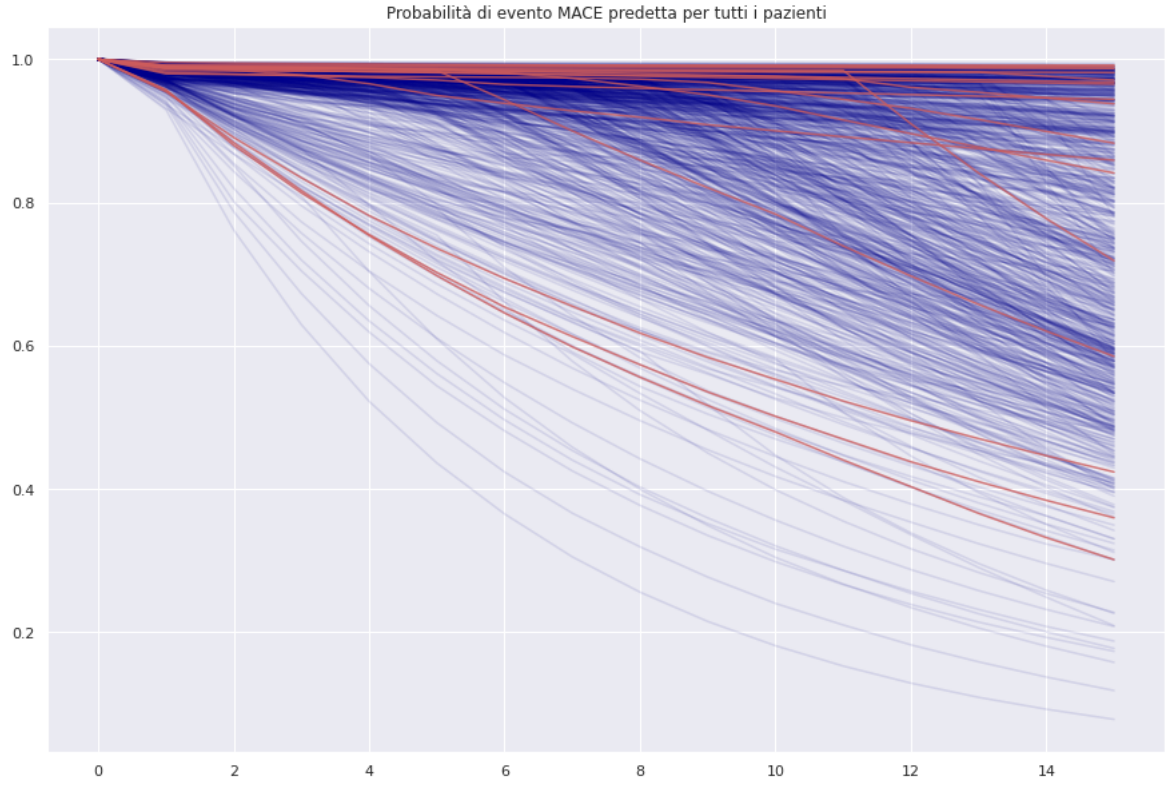

Figure 8.6: Predictive curves of the Dynamic DeepHit. Censored data (blue) vs patients with an event (red).

9 | Models with bootstrapped data

The data used in this work are very unbalanced, as already said only 2% of the patients have had an event, and the numerosity is reduced. Therefore the training process is difficult because a low number of events means low information available. The evaluation on the test set, which consists in only 20% of the data, is even more difficult, less robust and in particular is highly dependent on the choice of the test set. In order to test and evaluate the methodologies in our case study, we replicate the analysis on an augmented dataset. Since we are not able to encounter new data of the same type from other sources, we seek the solution with mathematical tools, by bootstrapping data and adding a white noise error. [12]
The bootstrap method is a technique for estimating quantities about a population by averaging estimates from small data samples. The sampling with replacement, i.e. after an observation is sampled it is returned to the possible sampling observations, is the key part of the process. "*The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method*" [12].

Here the bootstrap method is used in a slightly different way: we are going to sample from our data a bigger set. We are increasing the data dimension preserving its distribution. Such a process is not helpful without an additional error. Indeed without a white noise, i.e. a random error with zero mean and small variance, a model would interpolate the original data excessively and probably would underperforms on the test set. The Jittering process adds a small random error to each observation on both the covariates and the target variable. In this study the error was randomly sampled from a gaussian distribution with zero mean and with variance equal to a fourth of the feature's variance:

$$\epsilon_{ip} \sim \mathcal{N}(0, \sigma(\mathbb{X}_p)/4) \tag{9.1}$$

We have built a dataset seven times bigger than the starting one, we have selected seven because it was the higher value that the method was able to handle. We have tested two different models on the new dataset: the reduced Cox model and the reduced DeepHit model at the baseline with the covariates selected in section 7.4. Some metrics of the Cox model trained over the bootstrapped data are lower than the first two Cox models but others are higher such as the accuracy. The C-index grows on the training set (0.827) but decreases when measured on the

test set (0.623), moreover the accuracy (0.364) and the specifivity (0.354) grows a little but the sensitivity decreases (0.806). Instead the DeepHit model performs well on the bootstrapped data, in fact it reaches the lowest values in the mean squared error metrics (26.272 for the full model and 39.836 for the reduced one). In terms of C-index it has the lowest value on the training set (0.691) but it remains high on the test set (0.670) suggesting that it is a robust method. We conclude that the bootstrap process helps the training of neural network method whilst the Cox model does not improve its performances. All the metrics are reported in Tables 9.1 and 9.2. Due to computational problem we were unable to train the Dynamic DeepHit on the bootstrapped dataset, in fact the original longitudinal dataset is too big to perform bootstrapping.

Model	Training C-index	Test C-index
Cox	0.826	0.6603
Cox reduced	0.794	0.6912
Cox with bootstrap	0.827	0.6231
DeepHit	0.7113	0.7394
DeepHit reduced	0.8001	0.6564
DeepHit with bootstrap	0.6909	0.6701

Table 9.1: C-index for all the models at the baseline.

Model	Mse event	Mse all	Accuracy	Sensibility	Specificity
Cox	28.286	39.294	0.351	0.944	0.339
Cox reduced	21.285	36.373	0.280	1.000	0.266
Cox with bootstrap	37.587	42.300	0.364	0.806	0.354
DeepHit	41.606	47.231	0.069	0.944	0.051
DeepHit reduced	40.598	57.005	0.442	0.722	56.842
DeepHit with bootstrap	26.287	39.836	0.258	0.921	0.245

Table 9.2: Metrics adopted for the evaluation of the models at baseline.

10 | Conclusion

In 1998 the introduction of a new combination of inhibitor drugs into AntiRetroviral Therapies (ARTs) enabled people affected by Human Immunodeficiency Virus (HIV) to lead longer and healthier lives, but the possibility that these drugs increase the risk of CardioVascular Diseases (CVDs) has developed rapidly. Many studies analysed the relationship between some of these inhibitor drugs and the time to a CVD event. Until today, though, there still exists uncertainty about these relationships. Most of these studies sought short term relationships between ARTs and the time of CVD events and did not consider the newest ARTs developed after 2007. The need of a contemporary long-term analysis of the problem led the infectious diseases department of the Ospedale San Raffaele to collect data from the follow up of 4512 patients affected by HIV on a time window of 15 years. These data have been collected from 1998 until now, therefore they include the information about the integrase inhibitor (INIs) drugs that were introduced in the ARTs after 2007 and therefore have been poorly studied.

The majority of previous studies has approached this problem with classical regression and classification methods, and a few survival analysis tools. Recently, with the advent of sophisticated machine learning methods and great computational power, new methods have been developed to analyse the time to CVD events. In this work we analyzed the time to CVD events in patients affected by HIV within 15 years from the beginning of the ART. We compared the classical Cox PH model with the DeepHit model, that is a neural network approach for survival analysis, in terms of interpretability of the results and predictive power. Both methods are fitted at the baseline, i.e. considering personal and clinical information of patients only at the beginning of the ART, and, later, by including time-dependent covariates. Considering longitudinal data allows to track the patients' clinical history through time, in terms of illness diagnosis, clinical measurements and time of exposure to ART drugs.

The Cox models are easily readable, in fact estimating the hazard ratios we can interpret the behaviour of each covariate on the target variable, moreover they provide the quantitative significance of each feature. The DeepHit models are, per se, not interpretable, but, by applying Permutation Feature Importance (PFI) and Shapley value techniques, they provide a representation of the most important variables and of their relationship with time-to-event. It is worth to mention that these techniques provide a different qualitative approach that can explain well the non-linear relationships. Moreover, we are able to analyse also the interaction between two

different covariates and their impact on the target variable. Besides the rigid functional form, the main drawback of Cox models is the PH (PH) assumption that constrains its application. In this work the full Cox model has violated such assumption, therefore its results are not reliable. Removing the proper features we managed to fit a simple and robust model that satisfied the PH assumption. The DeepHit model does not assume the PH assumption and therefore it is more flexible. Moreover it is able to capture non-linear and even non-proportional interactions between covariates and time-to-event.

Of all the models at the baseline, the full DeepHit model has the highest C-index on the test set, it is therefore the best model with fixed-time variables if we want to predict correctly the ranking of patients. It is worth to underline that the difference between the C-indexes, measured on the training set and on the test set, are higher for the two Cox models with respect to the neural network models. We conclude that the two techniques used in this work, *dropout* and *weight regularization*, are able to prevent overfitting in neural network models.

With the introduction of longitudinal data into the analysis the models are able to capture time-dependent patterns between the covariates and the time-to-event. The two approaches gain better performances, in particular, the full Dynamic DeepHit model reaches the highest C-index on the test set. In terms of the other metrics, the time-dependent Cox models are more balanced, in fact, they make lower error in the prediction of the time to CVD events and are more precise in predicting those patients that had experienced an event.
Moreover, applying neural network methods to longitudinal data has increased the computational cost and the complexity, therefore the training process with time-dependent variables is longer and more difficult. The interpretability of this method is limited since some of the techniques used to explain the model cannot be applied.

The results of the models agree with clinical theory. In fact, the age of the patient is the variable that affects the most the risk of CVD events. The most significant variables identified by the models are hypertension, hepatitis C and the number of platelets, also the importance of these variables is clinically supported.
Moreover we have interesting results regarding the ARTs inhibitor drugs influence on the risk of CVD events. From the results of the Cox models we conclude that, with a significance of 0.4%, one year of exposure to NRTIs drugs decreases the risk of CVD events of more than 15%. Visualizing the Shapley values of the prediction of DeepHit we conclude that a very low time of exposure to NRTIs and PIs drugs increases the risk of CVD events, while longer exposure to these drugs are actually a protective factor. Moreover we have discovered that the response to these drugs is different between old and young patients. In fact for old patient the short exposure leads to higher risks with respect to young patients, on the other hand a long exposure is more effective for old patients than for young individuals that in this case have higher risk.

Even though the scarcity of available data, and in particular its imbalance, has limited the performances of all methods, this work has pointed out that in long term analysis the influence

of ART drugs decreases the risk of CVD events.
This first study leaves the basis to further analysis, in particular, the DeepHit model is able to explain well the non-linear relationships of interest and performs even better than the Cox model. The time-dependent data have increased the computation cost, but has allowed the models to reach good performances. In particular the Dynamic DeepHit, with great computational power, allows to analyse the relationship between covariates and time to CVD events as it varies over time. A further analysis, with more available data, may be conducted considering the different competing risks of CVD events to analyse in detail the influence of each inhibitor on each particular CVD event.

Bibliography

[1] What are hiv and aids. URL `https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids`.

[2] Hiv/aids. URL `https://www.who.int/news-room/fact-sheets/detail/hiv-aids`.

[3] Feature importances with a forest of trees. URL `https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html`.

[4] S. Ameri. Survival analysis approach for early prediction of student dropout. *Aggiungi: "Master thesis in computer science at the Graduate School of Wayne State University, Detroit, Michigan"*, 2015.

[5] C. Bavinger, E. Bendavid, K. Niehaus, R. A. Olshen, I. Olkin, V. Sundaram, N. Wein, M. Holodniy, N. Hou, D. K. Owens, et al. Risk of cardiovascular disease from antiretroviral therapy for hiv: a systematic review. *PloS one*, 8(3):e59551, 2013.

[6] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[7] J. S. Currier. Management of cardiovascular risk (including dyslipidemia) in patients with hiv. *UpToDate, Waltham, MA (Accessed on January 15, 2020.)*, 2020.

[8] F. D'Ascenzo, O. De Filippo, G. Gallone, G. Mittone, M. A. Deriu, M. Iannaccone, A. Ariza-Solé, C. Liebetrau, S. Manzano-Fernández, G. Quadri, et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (praise): a modelling study of pooled datasets. *The Lancet*, 397(10270):199–207, 2021.

[9] A. Di Castelnuovo, M. Bonaccio, S. Costanzo, A. Gialluisi, A. Antinori, N. Berselli, L. Blandi, R. Bruno, R. Cauda, G. Guaraldi, et al. Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with covid-19: survival analysis and machine learning-based findings from the multicentre italian corist study. *Nutrition, Metabolism and Cardiovascular Diseases*, 30(11):1899–1913, 2020.

[10] D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

[11] D. A. D. S. Group et al. Use of nucleoside reverse transcriptase inhibitors and risk of

myocardial infarction in hiv-infected patients enrolled in the d: A: D study: a multi-cohort collaboration. *The Lancet*, 371(9622):1417–1426, 2008.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] D. G. Kleinbaum and M. Klein. *Survival analysis*, volume 3. Springer, 2010.

[15] D. G. Kleinbaum and M. Klein. *Survival Analysis*. Springer, 3 edition, 2015.

[16] Knuth and D. E. Explain any models with the shap values: use of the kernelexplainer, 2019. URL `https://towardsdatascience.com/explain-any-models-with-the-shap-values-use-the-kernelexplainer-79de9464897a`.

[17] S. Lang, M. Mary-Krause, L. Cotte, J. Gilquin, M. Partisani, A. Simon, F. Boccara, D. Costagliola, C. E. G. of the French Hospital Database on HIV, et al. Impact of individual antiretroviral drugs on the risk of myocardial infarction in human immunodeficiency virus–infected patients: a case-control study nested within the french hospital database on hiv anrs cohort co4. *Archives of internal medicine*, 170(14):1228–1238, 2010.

[18] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Conference on Artificial Intelligence (AAAI)*, 2018. URL `http://medianetlab.ee.ucla.edu/papers/AAAI_2018_DeepHit`.

[19] C. Lee, J. Yoon, and M. van der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering (TBME)*, 2020. URL `https://ieeexplore.ieee.org/document/8681104`.

[20] J. L. Marcus, L. B. Hurley, D. S. Krakower, S. Alexeeff, M. J. Silverberg, and J. E. Volk. Use of electronic health record data and machine learning to identify candidates for hiv pre-exposure prophylaxis: a modelling study. *The lancet HIV*, 6(10):e688–e695, 2019.

[21] C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[22] V. Pillay-van Wyk, W. Msemburi, R. Dorrington, R. Laubscher, P. Groenewald, and D. Bradshaw. Hiv/aids mortality trends pre and post art for 1997-2012 in south africa–have we turned the tide? *South African Medical Journal*, 109(11b):41–44, 2019.

[23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

[24] J. A. Roth, G. Radevski, C. Marzolini, A. Rauch, H. F. Günthard, R. D. Kouyos, C. A. Fux, A. U. Scherrer, A. Calmy, M. Cavassini, et al. Cohort-derived machine learning

models for individual prediction of chronic kidney disease in people living with human immunodeficiency virus: A prospective multicenter cohort study. *The Journal of infectious diseases*, 224(7):1198–1208, 2021.

[25] S. Safo, L. Haine, J. Baker, C. Reilly, D. Duprez, J. Neaton, J. Wang, M. K. Jain, A. A. Pinto, T. Staub, et al. 89976 assessing protein biomarkers role in cvd risk prediction in persons living with hiv (pwh). *Journal of Clinical and Translational Science*, 5(s1):47–48, 2021.

[26] A. Scaccia. Facts about hiv: Life expectancy and long-term outlook. *Circulation*, 2020.

[27] M. Spreafico. Cox proportional hazard model with clinical application. *Biostatistics course. Politecnico di Milano*, 2019.

[28] V. A. Triant, J. Perez, S. Regan, J. M. Massaro, J. B. Meigs, S. K. Grinspoon, and R. B. D'Agostino Sr. Cardiovascular risk prediction functions underestimate risk in hiv infection. *Circulation*, 137(21):2203–2214, 2018.

[29] A. Tseng, J. Seet, and E. J. Phillips. The evolution of three decades of antiretroviral therapy: challenges, triumphs and the promise of the future. *British journal of clinical pharmacology*, 79(2):182–194, 2015.

[30] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[31] S. W. Worm, C. Sabin, R. Weber, P. Reiss, W. El-Sadr, F. Dabis, S. De Wit, M. Law, A. D. Monforte, N. Friis-Møller, et al. Risk of myocardial infarction in patients with hiv infection exposed to specific individual antiretroviral drugs from the 3 major drug classes: the data collection on adverse events of anti-hiv drugs (d: A: D) study. *The Journal of infectious diseases*, 201(3):318–330, 2010.

[32] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

A | Appendix A

A.1. KM estimator curves

Here we reported all the graphics of KM estimators for those variables that were not significant.

A.1.1. Independent-time covariates

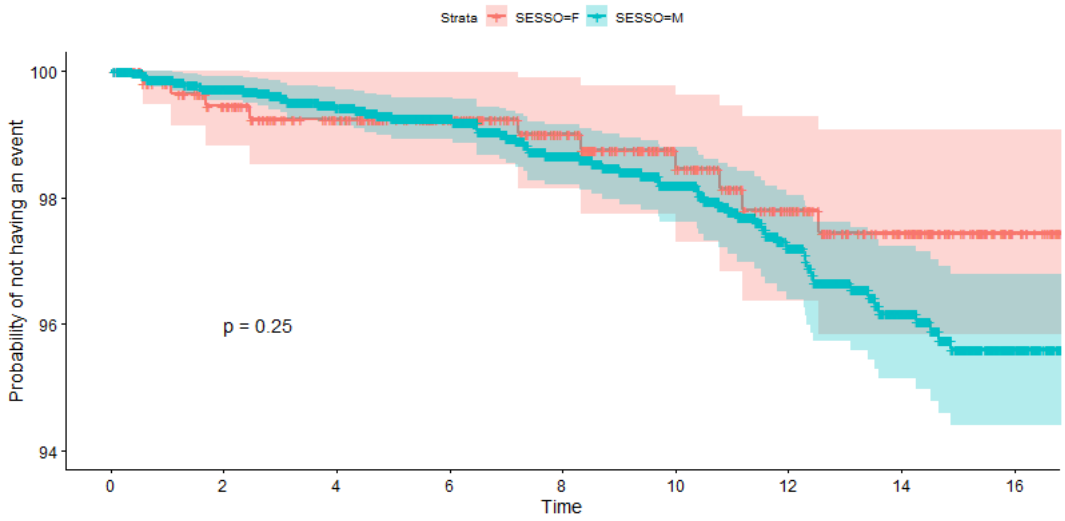

Figure A.1: KM estimator curves for sex. (1: Female ; 0: Male).

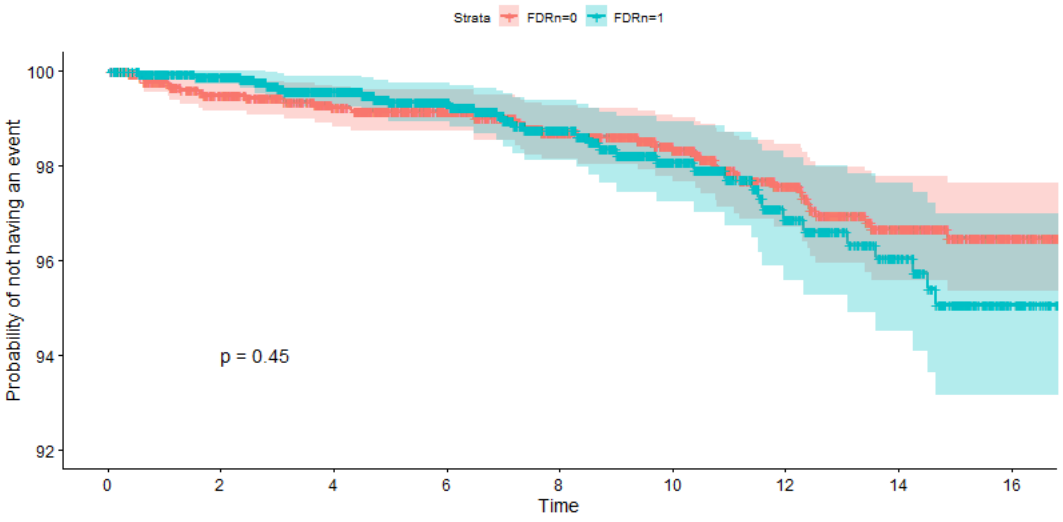

Figure A.2: KM estimator curves for factor of risk. (1: MSM; 0: Others).

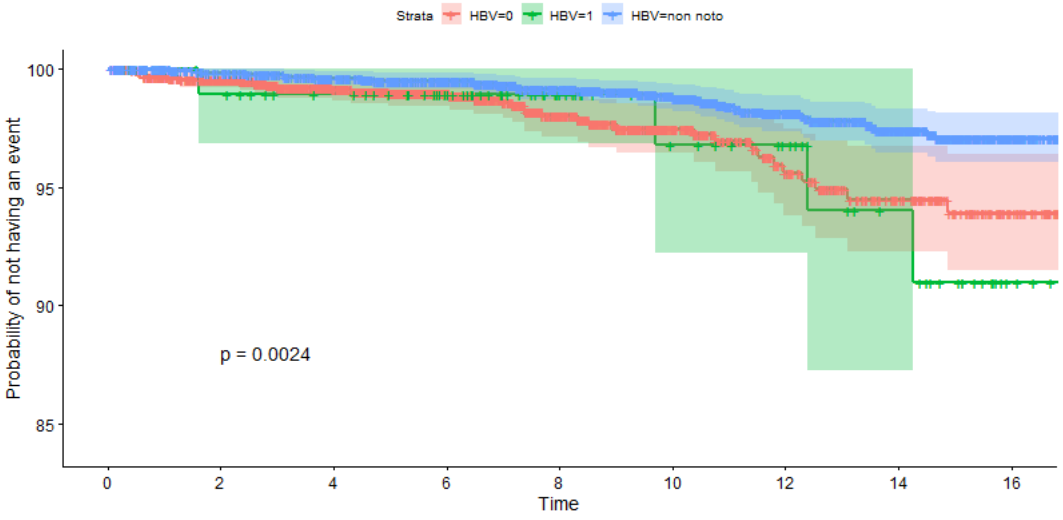

Figure A.3: KM estimator curves hepatitis B. (1: HBV present; 0: HBV absent; Not-known: HBV not known).

A.1.2. Time-dependent binary covariates

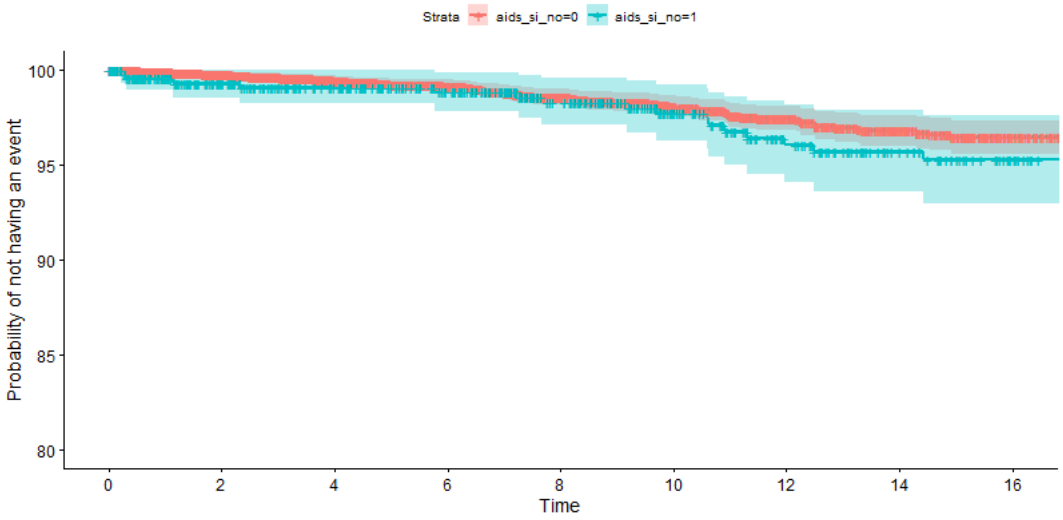

Figure A.4: KM estimator curves for the presence of aids. (1: Presence of AIDS; 0: Absence of AIDS).

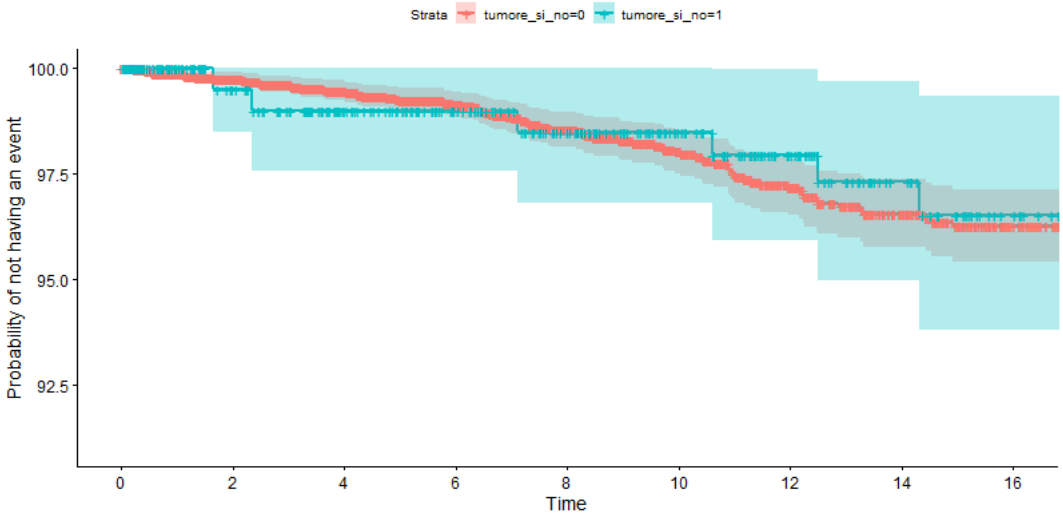

Figure A.5: KM estimator curves for the presence of tumor. (1: Presence of tumor; 0: Absence of tumor).

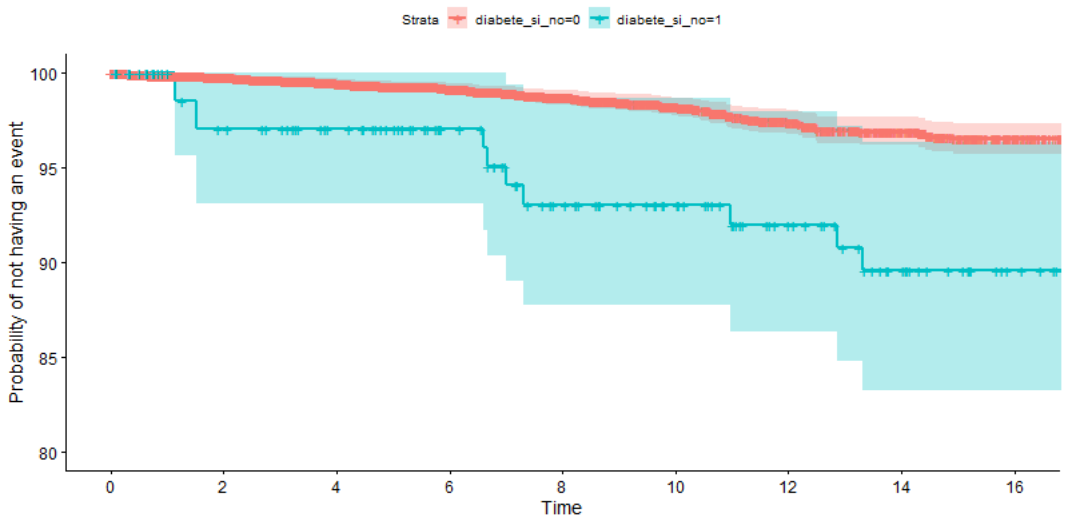

Figure A.6: KM estimator curves for the presence of diabetes. (1: Presence of diabetes; 0: Absence of diabetes).

A.1.3. Time-dependent numerical covariates

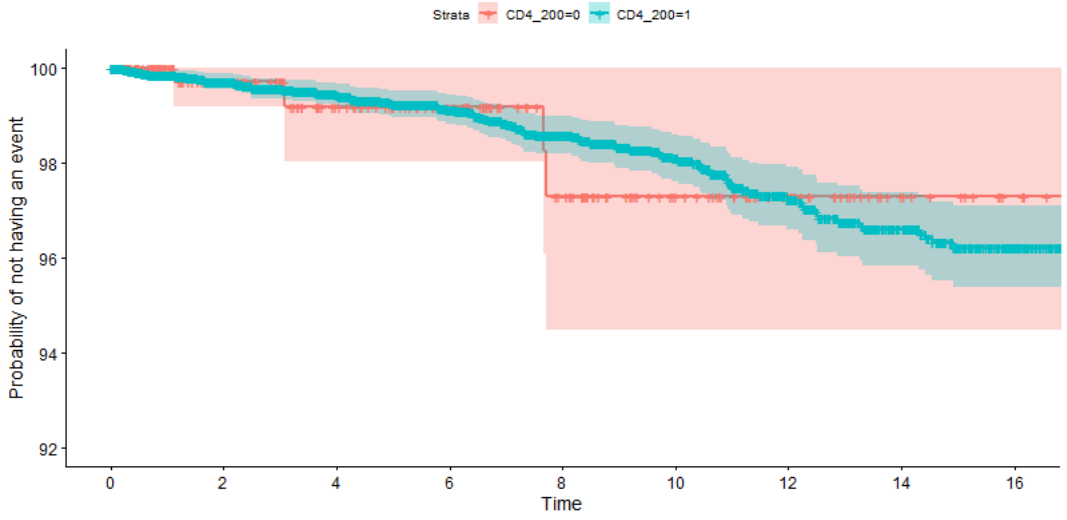

Figure A.7: KM estimator curves for the level of CD4 with cut-off at 200. (1: CD4 $\geq$ 200; 0: CD4 $<$ 200).

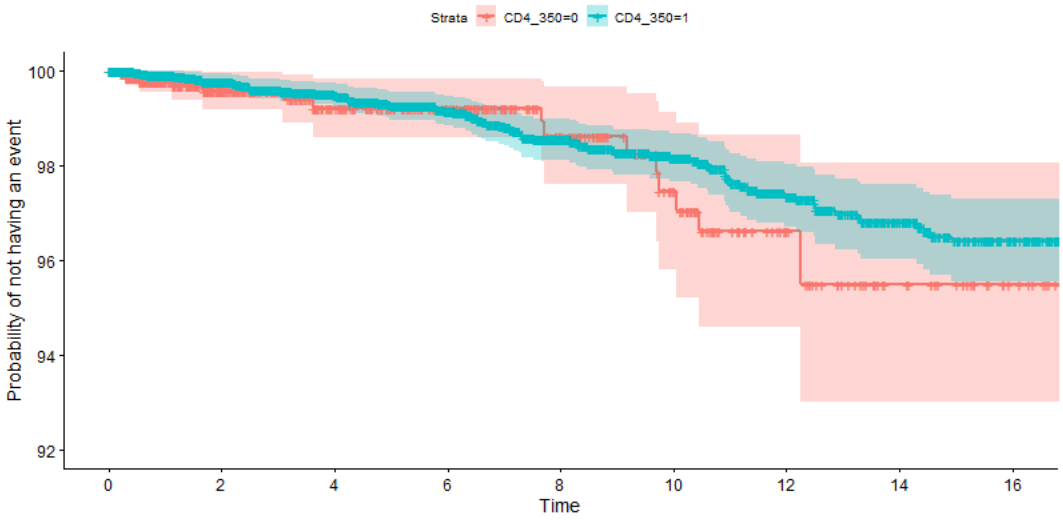

Figure A.8: KM estimator curves for the level of CD4 with cut-off at 350. (1: CD4 $\geq$ 350; 0: CD4 $<$ 350).

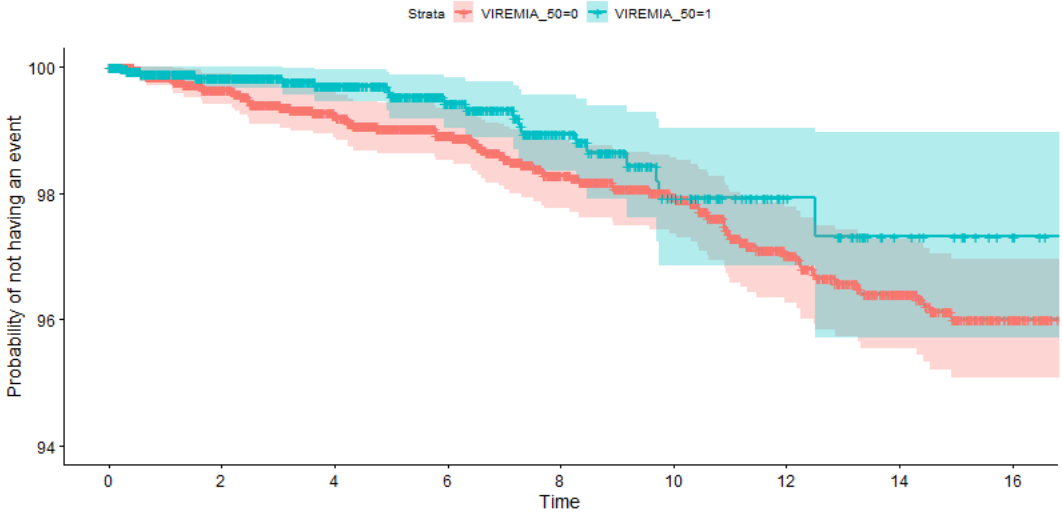

Figure A.9: KM estimator curves for the level of viremia (logarithm scale) with cut-off at 50. (1: Viremia $\geq log(50)$; 0: Viremia $< log(50)$).

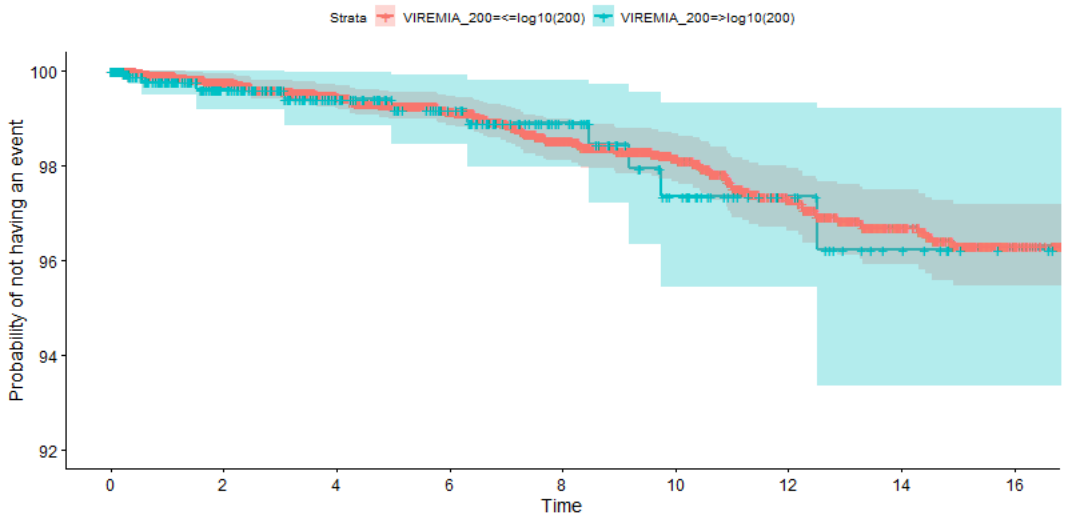

Figure A.10: KM estimator curves for the level of viremia (logarithm scale) with cut-off at 200. (1: Viremia $\geq log(200)$; 0: Viremia $< log(200)$).

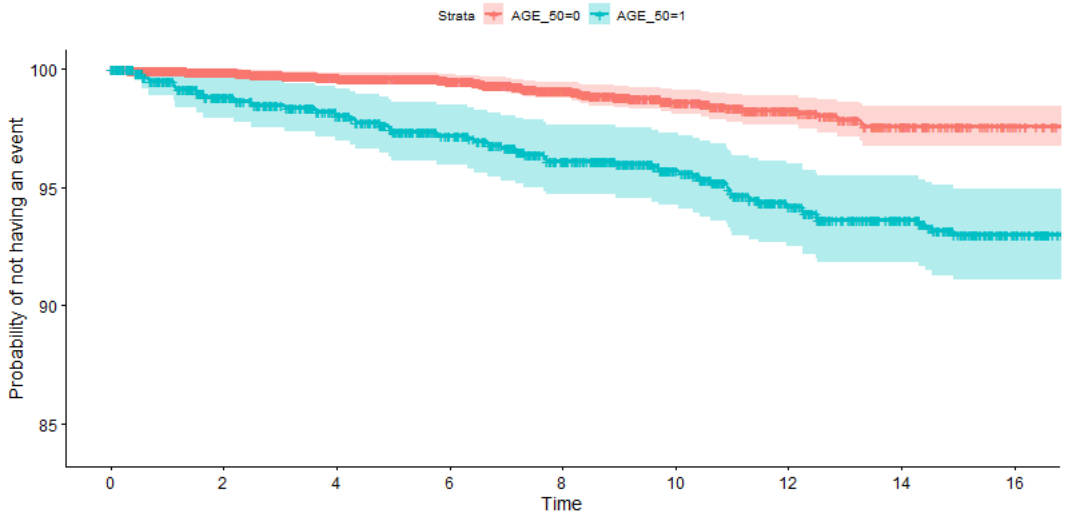

Figure A.11: KM estimator curves for the age of patients. (1: Age ≥ 50; 0: Age < 50).

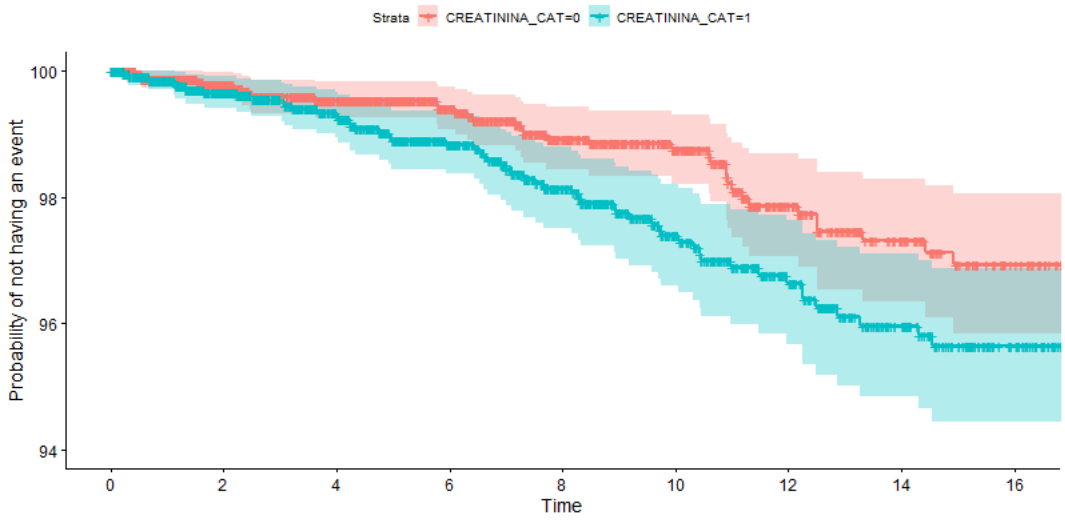

Figure A.12: KM estimator curves for the level of creatinine. (1: Creatinine $\geq$ median(Creatinine); 0: Creatinine $<$ median(Creatinine)).

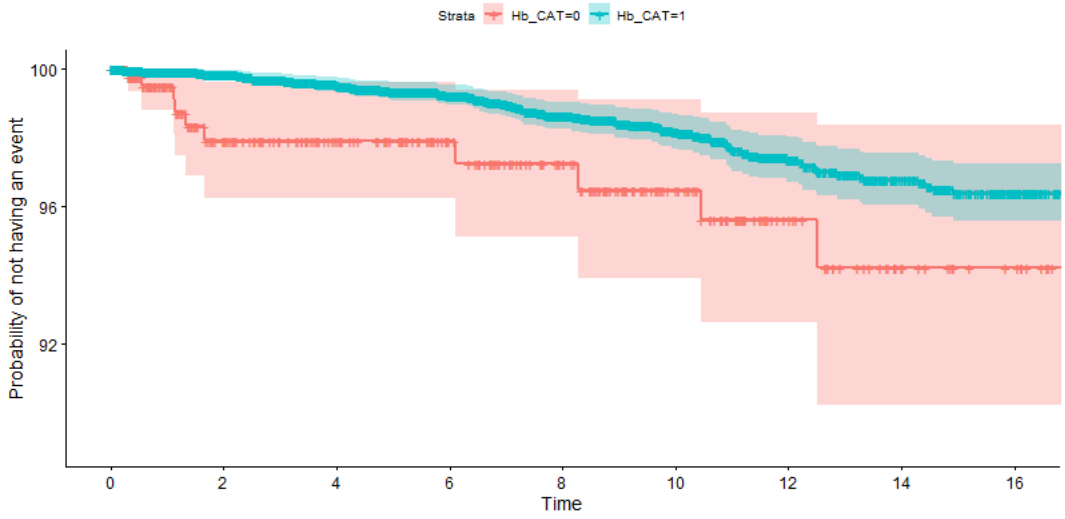

Figure A.13: KM estimator curves for the level of hemoglobin. (1: Hemoglobin $\geq$ 12; 0: Hemoglobin $<$ 12).

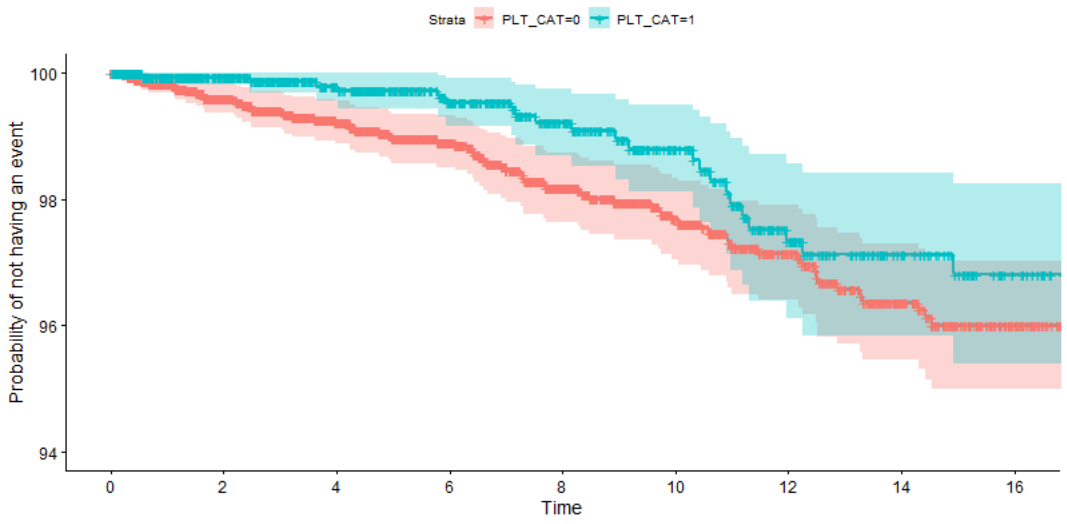

Figure A.14: KM estimator curves for platelets. (1: Platelets ≥ 250; 0: Platelets < 250).

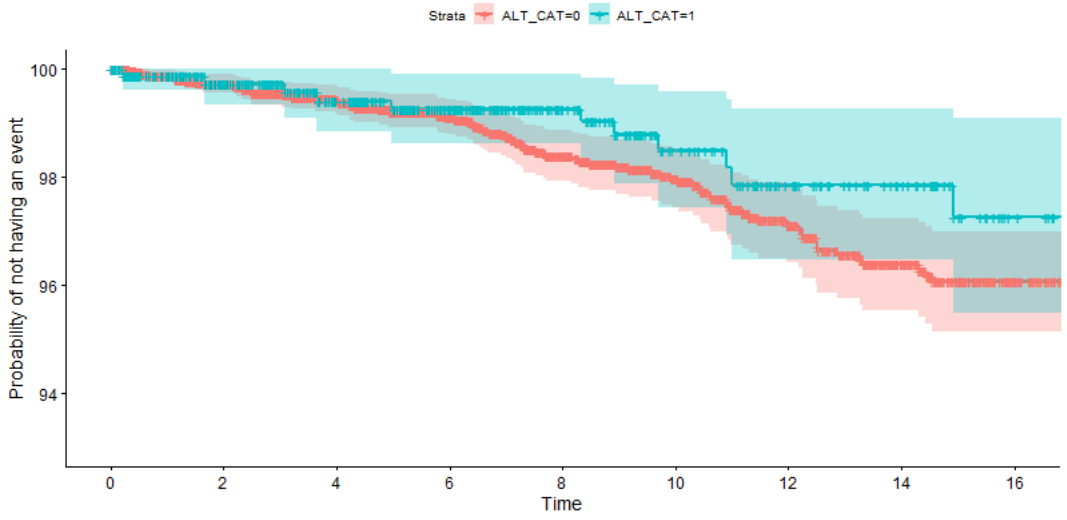

Figure A.15: KM estimator curves for the level of ALT with cut-off at 50. (1: ALT ≥ 50; 0: ALT < 50).

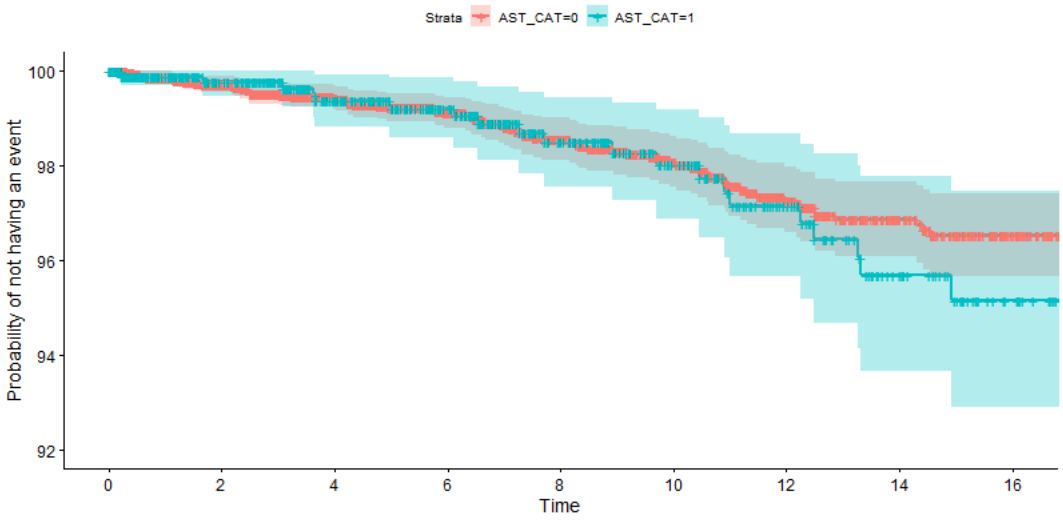

Figure A.16: KM estimator curves for the level of AST with cut-off at 35. (1: AST $\geq$ 35; 0: AST $<$ 35).

A.1.4. ARTs inhibitors

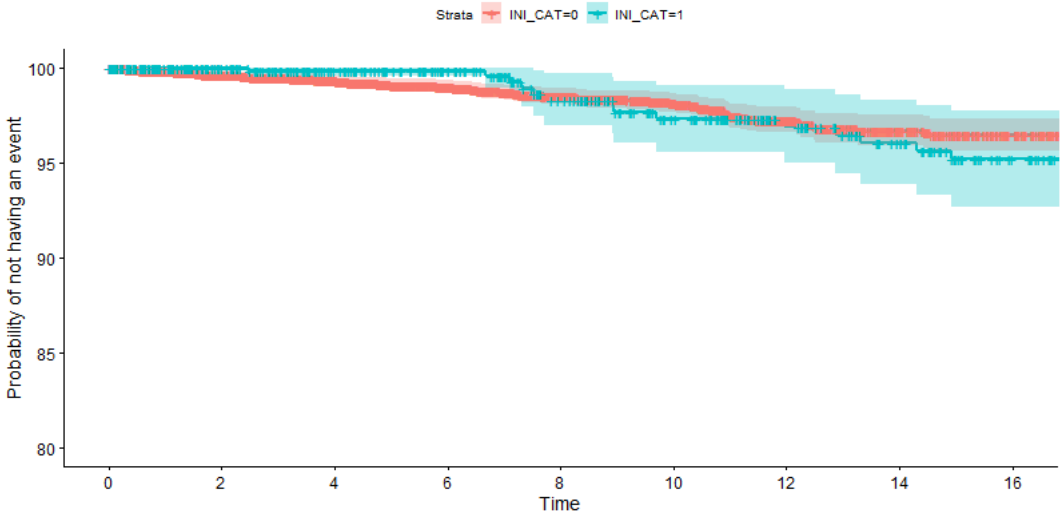

Figure A.17: KM estimator curves for the tie of exposure to INIs. (1: Exposure $\geq$ 6 months; 0: Exposure $<$ 6 months).

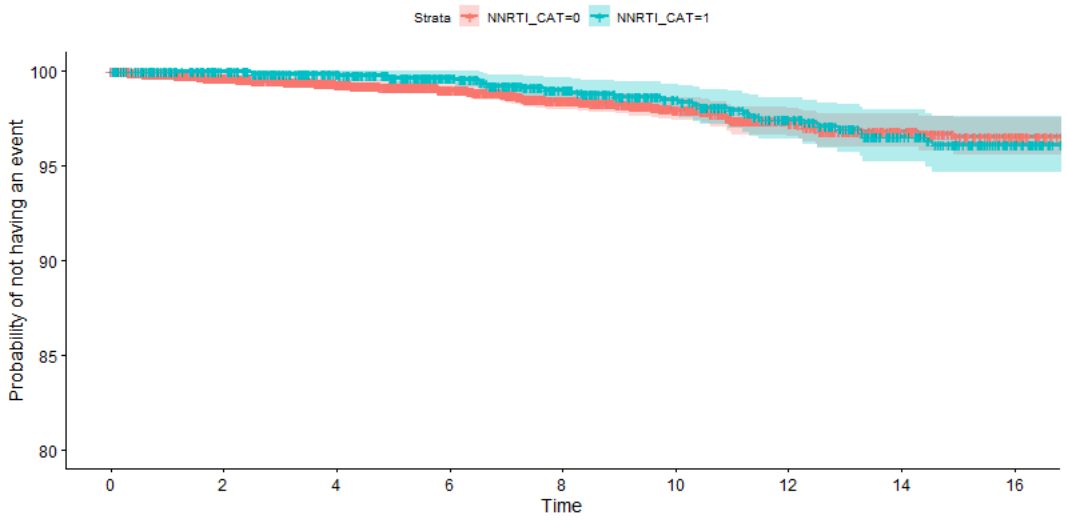

Figure A.18: KM estimator curves for the tie of exposure to NNRTIs. (1: Exposure $\geq$ 6 months; 0: Exposure $<$ 6 months).

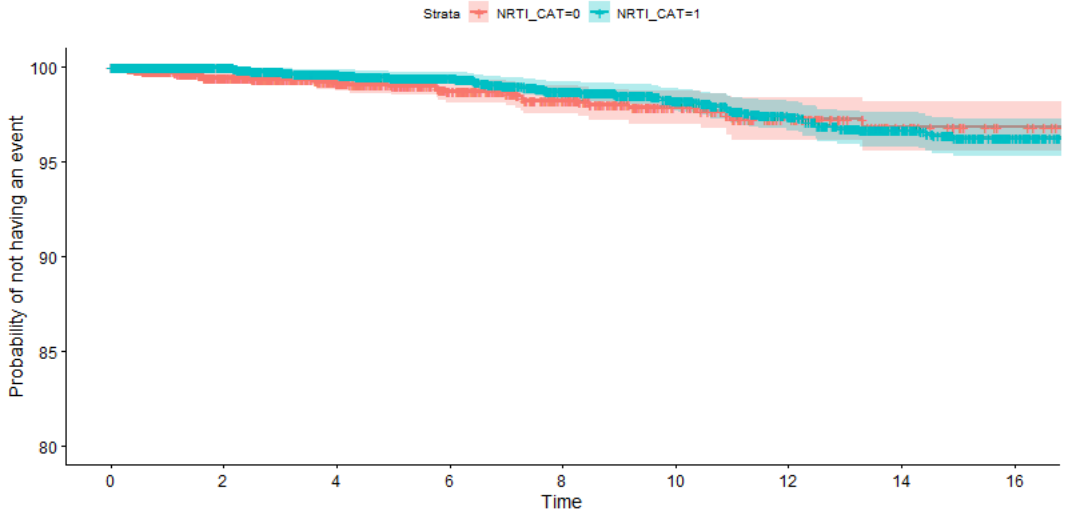

Figure A.19: KM estimator curves for the tie of exposure to NRTIs. (1: Exposure $\geq$ 6 months; 0: Exposure $<$ 6 months).

A.2. Data

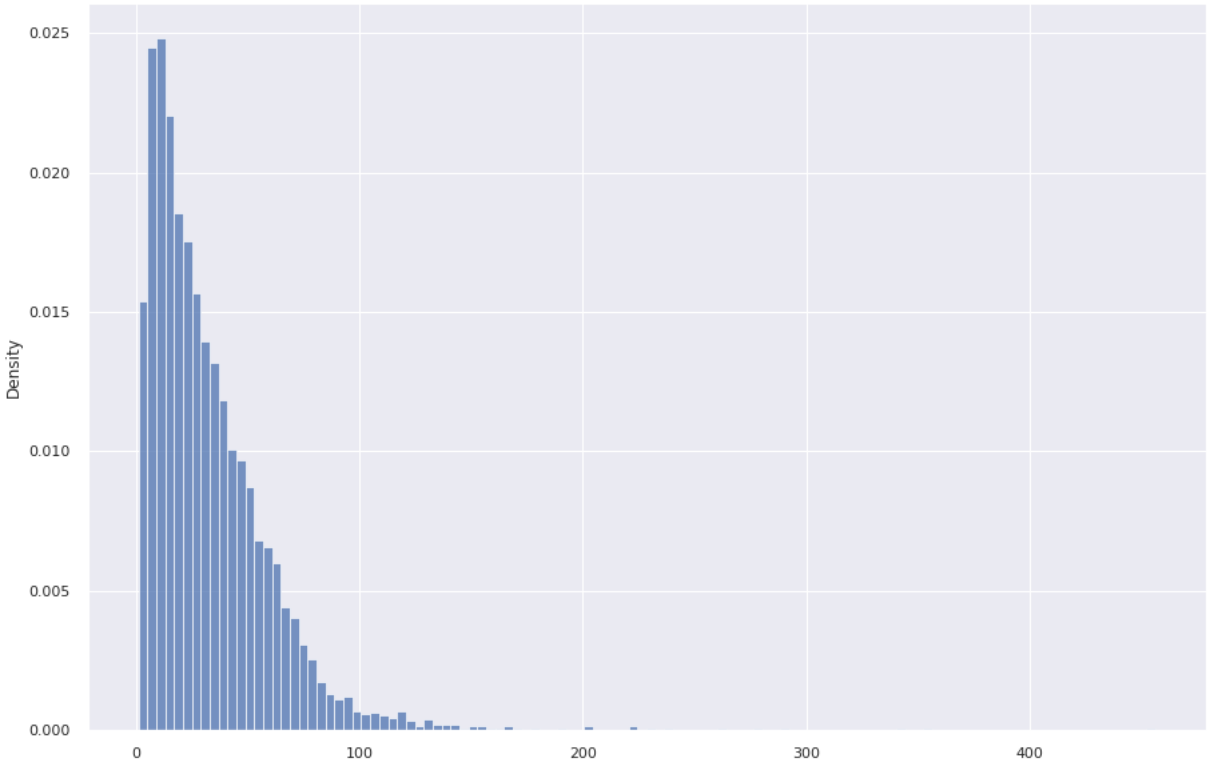

Figure A.20: Distribution of the number of visits for each patient.

B | Appendix B

Here are reported the diagnosis of the residuals of the Cox model [27] at the baseline, investigating the validity of the PH assumption for the full Cox model and the reduced one. We have reported three different plots of the residuals: Martingales residuals, deviance residuals and Schoenfeld residuals. The first two are for the whole model. Analysing the martingale residuals, the PH assumption is satisfied if the residuals have zero mean constant over time, x-axis. For the deviance residuals, that are a transformation of the martingales, the PH assumption is satisfied if the residuals are simmetric with respect to zero and with standard deviation equal to one. Here positive values correspond to individuals that had the event very soon compared to the estimated survival time while extreme values on the x-axis are outliers. Shoenfeld residuals represent the difference between the observed covariate and the expected given the risk set at that time. They should be randomly flat with zero mean.

B.1. Full Cox PH model diagnosis

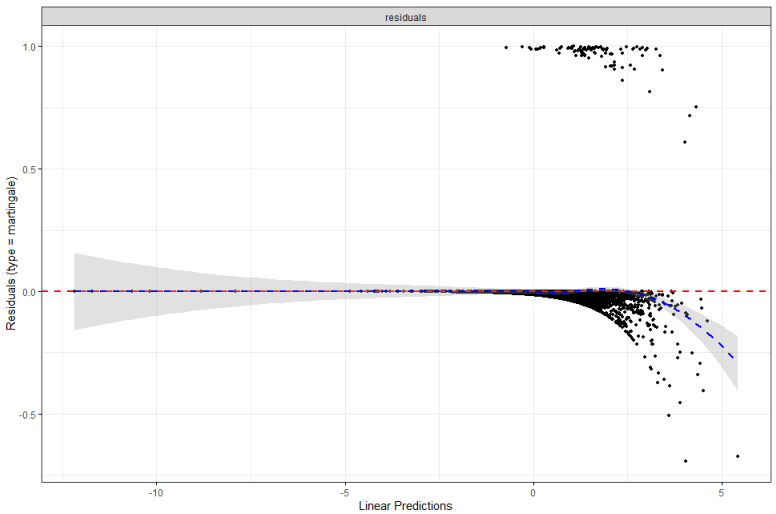

Figure B.1: Martingale residuals of the full Cox model.

Variables	chisq	df	p
Sex	$1.07e+00$	1	0.30016
Race	$1.90e-01$	1	0.66266
Year of ART	$1.60e+00$	1	0.20524
Factor of risk	$5.81e+00$	3	0.12131
HCV	$2.06e+00$	2	0.35756
HBV	$1.84e+00$	2	0.39769
Diabetes	$4.89e-01$	1	0.48439
Hypertension	$2.69e+00$	1	0.10070
Tumor	$5.16e-02$	1	0.82028
AIDS	$6.87e-06$	1	0.99791
CD4	$5.24e+00$	1	0.02214
Viremia	$5.18e-01$	1	0.47181
Age	$6.40e-01$	1	0.42380
Cholesterol	$2.15e+00$	1	0.14288
Hb	$2.82e+00$	1	0.09332
Platelets	$1.38e+00$	1	0.23930
triglycerides	$2.31e-01$	1	0.63065
Creatinine	$5.12e+00$	1	0.02371
ALT	$1.58e+00$	1	0.20891
AST	$4.62e-01$	1	0.49655
INIs exposure time	$1.69e+01$	1	4e-05
PIs exposure time	$7.32e-05$	1	0.99317
NRTIs exposure time	$1.00e+01$	1	0.00155
NNRTIs exposure time	$7.96e+00$	1	0.00477
GLOBAL	$6.13e+01$	28	0.00028

Table B.1: Test the proportional hazards assumption for each covariate of the full Cox model.

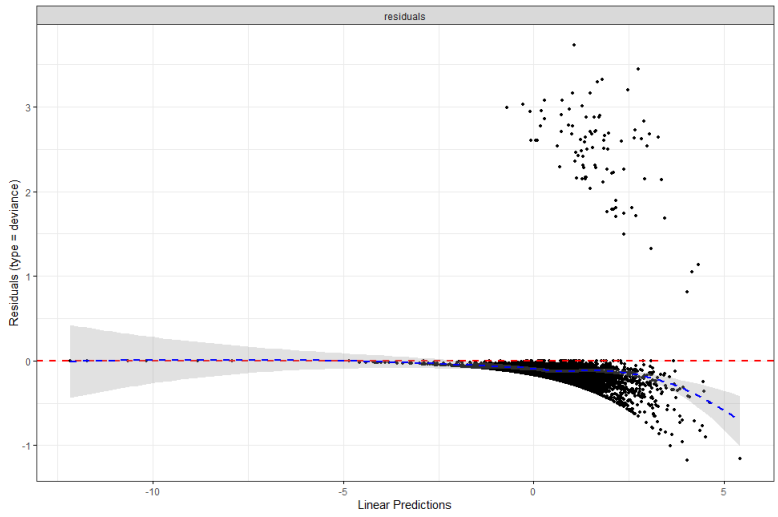

Figure B.2: Deviance residuals of the full Cox model.

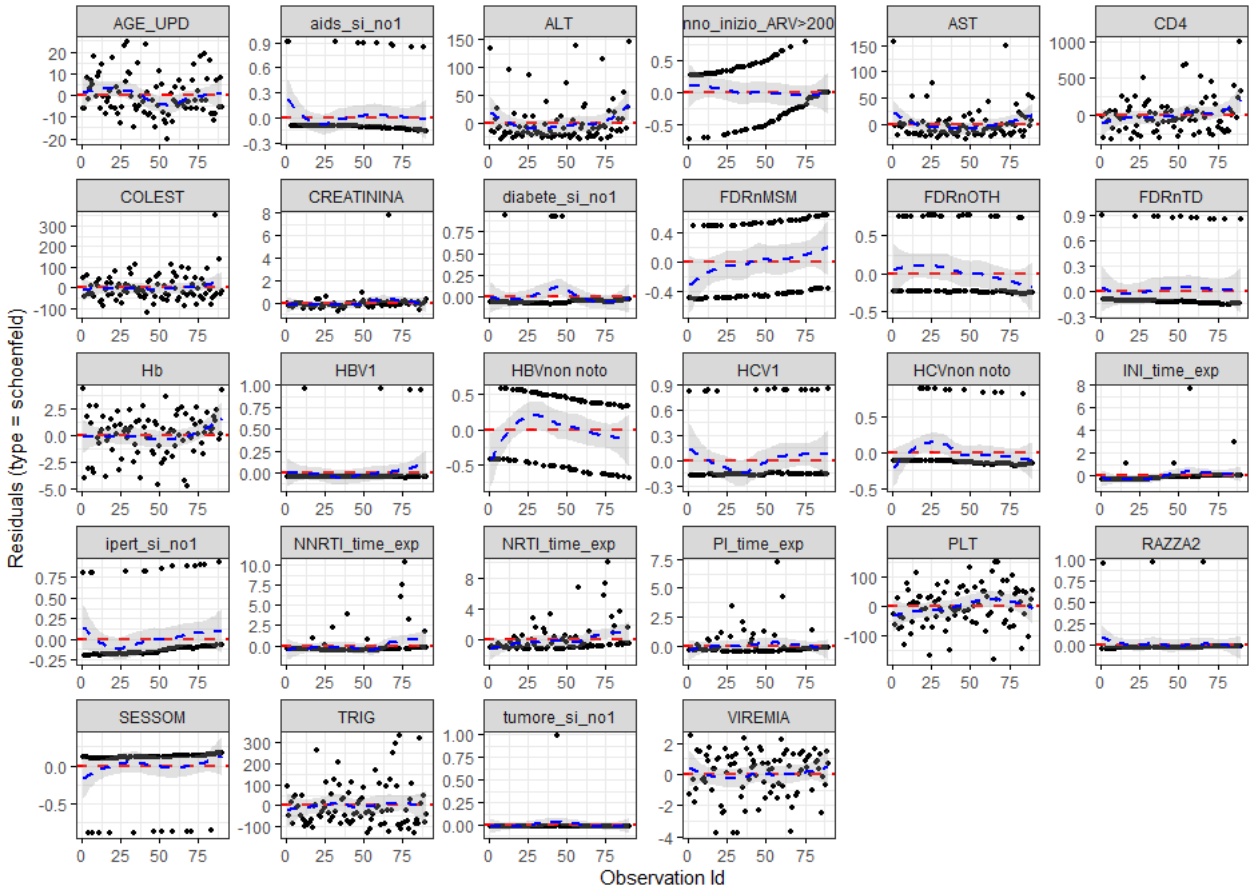

Figure B.3: Schoenfeld residuals of the full Cox model.

B.2. Reduced Cox PH model diagnosis

Variables	chisq	df	p
Year of ART	0.8998	1	0.34
HCV	0.3433	2	0.84
PIs exposure time	0.0475	1	0.83
Hypertension	1.8921	1	0.17
Diabetes	0.4268	1	0.51
Age	0.3204	1	0.57
Creatinine	2.6284	1	0.10
GLOBAL	8.3426	8	0.40

Table B.2: Test the proportional hazards assumption for each covariate of the reduced Cox model.

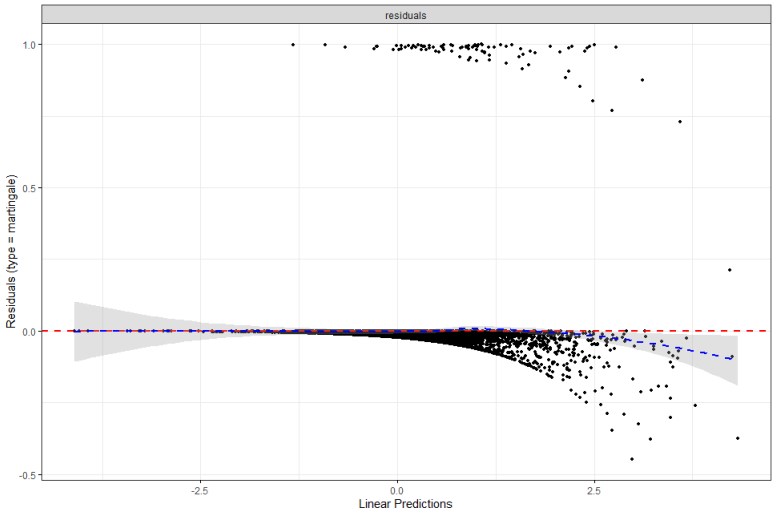

Figure B.4: Martingale residuals of the reduced Cox model.

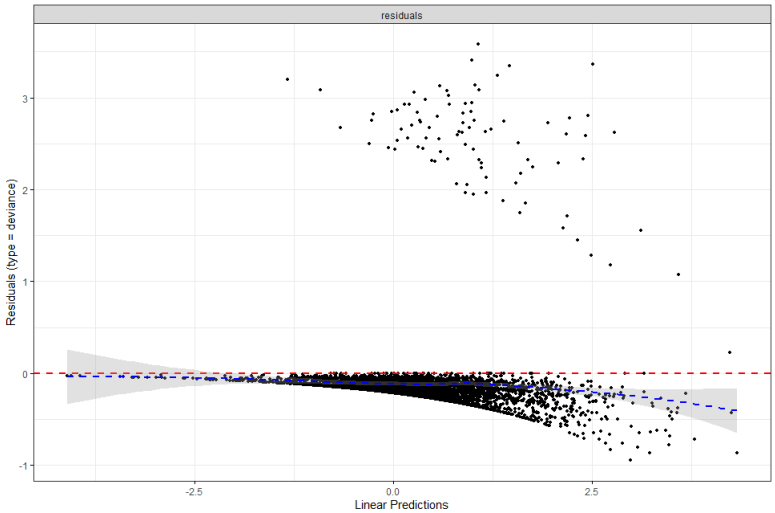

Figure B.5: Deviance residuals of the reduced Cox model.

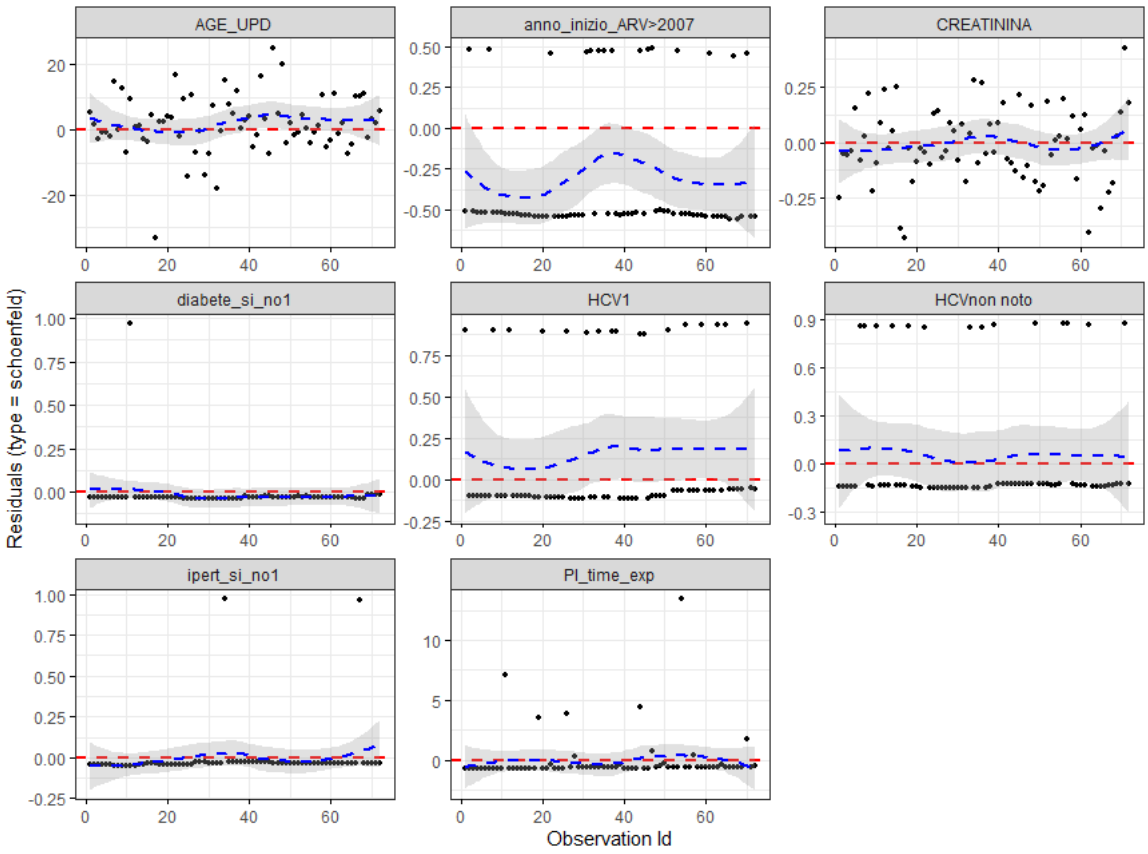

Figure B.6: Schoenfeld residuals of the reduced Cox model.

List of Figures

List of Tables

Acknowledgements

Ringrazio la professoressa Chiara Masci per la sua disponibilità e per avermi trasmesso una grande competenza, fondamentale per concludere con molta soddifazione questo lavoro. La ringrazio per la serenità con cui mi ha contagiato durante tutto questo percorso. Ringrazio l'ingegnere Federica Corso per il prezioso aiuto e per la gentilezza con cui mi ha supportato.