POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Mathematical Engineering - Statistical Learning

Author: Agostino Lurani Cernuschi

Advisor: Prof. Chiara Masci

Co-advisor: Ing. Federica Corso

Academic year: 2020-2021

1. Introduction

The Antiretroviral therapy, ART, has enabled individuals affected by human immunodeficiency virus, HIV, to live longer lives. In 1998 with the introduction of PIs [1] combined with NRTIs [2] and NNRTIs [3] the expected rate of death in people affected by HIV has dropped by 60 %. Recently the relation between these ART drugs and cardiovascular diseases, CVDs, has been investigated, in fact, even though the relations between each class of inhibitors and CVD are partially unclear, it appears that the exposure to some of these drugs increases the risk of CVD events [1]. Moreover the long-term relationship between INIs [4] and CVD events has not been studied since they were introduced in 2007. In this work we analyse the long-term relation between the time that passes between the start of the ART and the occurrence of a CVD event for patients affected by HIV with a time horizon of 15 years. We apply survival analysis tools on a real dataset, supplied by IRCCS Ospedale San Raffaele that has collected information from the follow-up of 4512 patients across the last 23 years. Firstly we analyse the time to CVD event in relation with data at the baseline, i.e. clinical and personal information at the beginning of the ART, comparing the results of two different methods: the classical survival analysis Proportional Hazard Cox model [2] and the neural networks based DeepHit model [6]. Then we include time-dependent covariates and we analyse how the relation between ARTs and CVD events varies over time applaying two other models, the Extended Cox model [4] and the Dynamic DeepHit [5], to longitudinal data. The goal of the thesis is therefore twofold: to analyse the relation between ARTs and the time to CVD events and to compare the classic survival analysis approach and the machine learning one that consists in a more flexible method that relaxes the linear assumption. In section 2 we introduce the basics of Survival analysis and the methods used, with the comparison of their points of strength and weakness. Then in section 3 we apply the methods to our dataset analysing the relation of interest and comparing the two different approaches. We draws conclutions in

[1] Protease Inhibitors
[2] Nucleoside Reverse Transcriptase Inhibitors
[3] Non-Nucleotide Reverse Transcriptase Inhibitors
[4] Integrase inhibitors

section 4

2. Methods

2.1. Classic survival analysis approach

Survival analysis is widely-used in many areas, such as economics and medicine, the variable of interest is the hitting time to a certain event, e.g. the time before an individual recovers from an illness or the lifetime of a bolt. The main issue in most survival analysis studies is the censoring problem, i.e. some observations are lost during the follow up. In such a case we have the information of an observation that has not experienced the event until a certain point but we do not know after the censoring time what may happen. We introduce the target variable as the couple of the survival time and the censoring indicator:

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

where:

- $T_i = min(T_i^*, C_i)$ is the survival time, T_i^* is the time to event and C_i is the time to censoring of the i^{th} observation.
- $\delta_i = \mathbb{I}(T_i^* \leq C_i)$ is the indicator random variable that indicates whether an i^{th} observation is censored or not.

The *Survival function* $S(t)$ is introduced to describe the probability of non having an event until each time t:

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t) = 1 - F(t) \tag{2}$$

The instanteneus risk of event is described by the *Hazard function* that is the limit for Δt that goes to 0 of the probability of having the event in such time interval, having survived until time t:

$$h(t) = \mathbb{P}(T = t | T \geq t) \tag{3}$$

2.1.1. The Kaplan-Meier estimator

The Kaplan-Meier [4] is a univariate estimator that approximates the Survival probability considering the number of individuals that experience the event at time t over the number of individuals at risk, i.e. that did not have the event nor the censoring before time t. If we consider two different groups of individuals we can test, through the log-rank test if two *survival functions* are significantly different. From the survival curves of each group it is possible to get a qualitative interpretation of which group has lower probability of not having an event.

2.1.2. The Proportional Hazard Cox model

The Proportional Hazard Cox model is a semi-parametric model that can associate covariates with the time-to-event. The model hazard function for an individual i is defined as:

$$h_i(t|\mathbb{X}_i) = h_0(t)e^{\mathbb{X}_i^T \beta} \tag{4}$$

where $h_0(t)$ is the baseline hazard function that does not depend on the covariates, $\mathbb{X}_i^T \beta$ is the product between the covariates of the i^{th} individual and the vector of coefficients. The estimate of the vector of parameters is made through the maximum likelihood estimation of the *Partial likelihood*, i.e. a likelihood that considers only the probabilities for all patients that experience the event. The Cox model makes some strong assumptions on the underlying stochastic process: the hazard ratios must be constant over time and the relationship between log hazard and covariates must be linear.

2.1.3. The Extended Cox model

If we consider longitudinal data we can still use the Cox model, but, since it no longer satisfies the proportional hazard assumption, we need to use an extended version: the Time-Dependent Cox model [4]. In fact, here the baseline hazard and survival function depend on time, but the hazard at time t depends only on the value of the time-dependent covariates at the same time t. The model is defined as before, but the baseline survival function depends on time:

$$h_i(t|\mathbb{X}(t)) = h_0(t)e^{\left[\sum_{k=1}^{P_{fix}} \mathbb{X}_{ik}\beta_k + \sum_{z=1}^{P_{td}} \mathbb{X}_{iz}(t)\delta_z\right]} \tag{5}$$

where P_{td} is the number of time-dependent covariates and P_{fix} of those that are not, δ and β are the vectors of coefficients, respectively for time-dependent and time-fixed covariatses. It is worth to underline that the coefficients do not depend on time.

2.2. Neural network based approach to survival analysis

The second approach we present is based on a neural network. Neural networks are a method that in recent years has been widely used. It is based on

a high number of parameters connected with each other that are able to capture very complex patterns. To estimate these parameters they use an approximation of a loss function, usually through Stochastic Gradient Descent based methods. Each different problem needs the definition of a particular loss function that has the scope of minimizing the error of interest. The parameters of the net are divided in layers that are connected in sequence in feed-forward networks, and also have an auto connection loop to handle time-dependent data, in recurrent neural networks.

2.2.1. DeepHit

The first model we introduce will be trained on the baseline data. It is structured in different feed-forward block: a shared sub-network and, in case of multiple events, k cause-specific networks. The shared sub-network captures the patterns of the relationships between covariates and the time to event that are common to all different k events (in our application, k=1). Each cause specific network, that takes as input original data as well as the patterns that output from the shared block, captures the specific features of a particular event. It is based on the minimization of the weighted sum of two loss functions. The first is the log-likelihood of the joint distribution of the first hitting time modified to take into account right censored data. The second is the ranking loss that incorporates estimated cumulative incidence functions in order to finetune the network to each cause-specific event.

2.2.2. Dynamic DeepHit

In order to handle time-dependent data, the structure of the DeepHit is modified. In particular, we introduce a recurrent network that is able to capture time-dependencies. The structure of the total net is the same of before but the shared block is replaced with a shared recurrent neural network and the k specific networks do not receive directly the original data as input. Moreover introduce in the shared network a loss function that concentrates on time data and identifies the most important patterns of the time-dependent variables.

3. Application of the models to the dataset

In this section we introduce the data supplied by the Ospedale San Raffaele and then we analyse and evaluate the methods introduced previously applying them to both the baseline dataset and the longitudinal one.

3.1. HIV patients data

Data from 4512 patients affected by HIV were collected at IRCCS Ospedale San Raffaele from 1998 until today. Data previous to 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 have selected 24 variables from all those supplied by the IRCCS Ospedale San Raffaele. These variables were selected considering the proportion of missing data and consulting clinicians and doctors. 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: NNRTIs, NRTIs, PIs and INIs. Moreover, we have added whether the patient has started the ART before or after 2007 since in that year the therapy was changed with the introduction of INIs. It is worth to note that among 4512 patients, only 90 experienced a CVD event within 15 years from the beginning of the ART.

3.2. Models at the baseline

The first models are fitted considering the data available at the baseline, i.e. the beginning of the ART. We analyse models trained on all the 24 covariates and then, after feature selection, we propose reduced models. We analyse the results and then we compare the models in terms of the following performances:

- C-index (training and test set);
- MSE of the predicted survival time evaluated on patients that had the event and on all the patients (test set);
- Accuracy in predicting patients that experienced an event (test set).

3.2.1. Full models

The full models are trained considering all the covariates.

Interpretation and results. The full Cox PH model results are shown in terms of hazard ratios and relative p-values. The age of the patient is the most significant variable followed by the presence of hepatitis C at the baseline, whether the patient is hypertensive, the level of creatinine and whether the patient has started the ART before or after 2007. In particular the last variable is the only protective factor, i.e. those patients who have started the ART before 2007 are less likely to experience a CVD while all others are risk factors. The information about the specific ART drugs were not significant, possibly because of the correlation they have between them-

selves. The full DeepHit can be interpreted through two techniques that are: Permutation Feature Importance, that shows the contribution of each covariate to the final prediction, and the Shapley value Additive Explanation, that shows both the importance of the variables and the dependency between each feature and the target variable. The most important variables are similar to the ones sought by the full Cox model, however the DeepHit found the ART drugs to be very important. In particular analysing the relationships between these covariates and the target variable we found out that the exposure to ART drugs is a protective factor. The full Cox model violated the PH assumption, therefore its results are not reliable.

Evaluation of the models. The models perform similarly in terms of all metrics, the Cox model is able to predict more accurately the patients that had experienced the CVD events and make a lower mean squared error in the prediction of the time to CVD event. The DeepHit model reached the higher C-index (0.74) on the test set and seemed to be more robust with respect to overfitting. All the metric performances are reported in Tables 1 and 2.

3.2.2. Reduced models

By a one-step procedure, with the help of the hazard ratio and the PH assumption, we build a new reduced Cox model with the most significant variables, that are: the years of exposure to NRTIs, the presence of hepatitis C and diabetes, whether the patient is hypertensive, the age and the level of creatinine. The reduced DeepHit model is trained on the following variables: the year of ART, the presence of hepatitis C, the Age, the number of Platelets, the level of creatinine and the years of exposure to PIs, NNRTIs and NRTIs (See Figure 1).

Interpretation and results. The reduced Cox PH model shows some new significant variables, in particular, in addition to the variables that resulted previously significant, the years of exposure to NRTIs, as well as the level of CD4, now results significant. Moreover this model does not violate the proportional hazard assumption and thus these results are reliable. The hazard ratios, reported in **??**, indicate that a year of exposure to NRTI reduces the risk of CVD events of more than 15%, with a level of significance of 0.4%. The DeepHit model highlights some interesting behaviour of the relationships between the covariates regarding the ARTs and the risk of CVD. In particular, the short exposure to NRTIs, x-axis, increases the risk of CVD events (higher values of Shapley indicate higher risks), but the medium-long exposure decreases this risk. Moreover, the interaction with the age of the patient indicates that the effect of the therapies is stronger in old patients both in increasing the risk, short exposure, and in decreasing it, long exposure (See Figure 2). Moreover also the PIs, the INIs and the NNRTIs result to be important variables for the model.

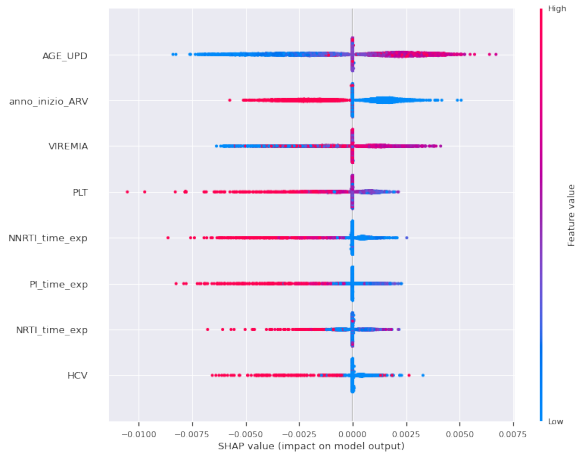

Figure 1: The Shapley value importance of the reduced DeepHit model at the baseline.

Evaluation of the models. Reducing the Cox model, that now does not violate the PH assumption, allows to trust the results, moreover it has similar performance as before and increases the C-index on the test set. The DeepHit reduced model has similar C-index with respect to the complete model and increases the accuracy of the predicted patients that had the event.

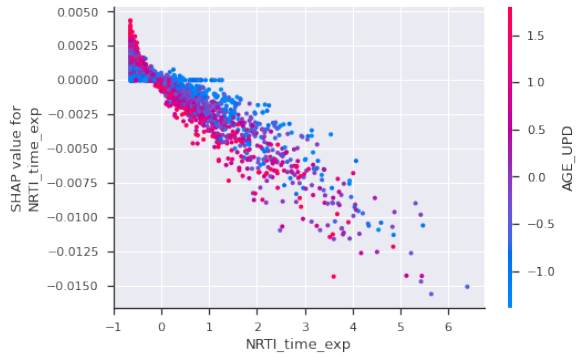

Figure 2: The Shapley value dependency on the time of exposure to NRTIs as the age of the patients varies for the reduced DeepHit model at the baseline.

3.3. Models with time-dependent variables

The introduction of longitudinal data, i.e. time dependent variables, adds information to the data, this makes the models more complex but also enables

to increase the performances. The most important thing though is that we are able to capture the relationship between a covariate and the time to CVD event as it varies over time and we can track the time to exposure to the ART drugs.

3.3.1. Full models

The full models were fitted on all the 24 covariates (some of them are fixed, others time-dependent); the Dynamic DeepHit was difficult to train for its complexity and for the fact that it is very time demanding, but it reaches interesting results.

Interpretation and results. The full models have highlighted some new variables that were not considered significant at the baseline, this suggests that the variation of this variable is important in the prediction of the risk of CVD events. The Time-Dependent Cox full model identifies, in addition to previous results, that the level of triglycerides, as well as the level of cholesterol, is very significant. The Dynamic DeepHit found the number of platelets to be the only important variable that was not already identified at the baseline. With time-dependent data we are unable to use the Shapley additive explanation due to the computational cost and therefore the interpretation of the results limited.

Evaluation of the models. The models increase their performances, in particular, the Dynamic DeepHit model reaches the highest C-index on the test set (0.77) and it is thus a good model to predict time to CVD events. In terms of the others metrics it preforms similar to the baseline DeepHit. The Cox model has a lower C-index on the test set but has higher accuracy than the neural network approach.

3.3.2. Reduced models

Here we analyse the two models on considering a reduced set of covariates in order to have clean, simpler and robust models. We perform feature selection using the hazard ratios 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. 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. 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.

Interpretation and results. The results are similar to the two complete models. The reduced time-Dependent Cox model indentifies that 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 results of the Dynamic DeepHit reduced model are similar to the complete model: the most important variables regard the ARTs inhibitors, in particular the NNRTIs and the PIs result to be the most important variables 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 the year of the beginning of the ART results as one of the most important.

Evaluation of the models. The Time-Dependent Cox with the reduced covariates has a higher value of the C-index on the test set, this suggests that the model is more robust towards new data. In terms of mean squared error, there are no significant changes, as well as for the accuracy and the sensitivity. The Dynamic DeepHit reduced model does not perform as well as the complete model. This suggests that all time-dependent covariates add some important information.

3.4. Models with bootstrapped data

Since the scarcity of data and the unbalances response have constrained the performances of all models we have decided to use a technique to augment data information. In particular we have bootstrapped the data [3], i.e. sampling with replacement, to obtain a dataset seven times bigger that has the same distribution of the original one. Then we have jittered the data, i.e. we added random noise, in order to have more information without repeating more times the same information. The results of this technique were not applicable to longitudinal data due to computational cost. Thus, we have trained the reduced baseline models on bootstrapped data, while the Cox model decreases its performances, especially in terms of C-index on the test set, the DeepHit model has improved its performances in almost all the metrics (See Tables 1 and 2). This shows how this neural network method is able to perform better with more data.

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
Extended Cox	0.809	0.6782
Extended Cox reduced	0.7998	0.7002
Dynamic DeepHit	0.7701	0.7713
Dynamic DeepHit reduced	0.7601	0.7228

Table 1: C-indexes of all the models

Model	Mse event	Mse all	Accuracy
Cox	28.286	39.294	0.351
Cox reduced	21.285	36.373	0.280
Cox with bootstrap	37.587	42.300	0.364
DeepHit	41.606	47.231	0.069
DeepHit reduced	40.598	57.005	0.442
DeepHit with bootstrap	26.287	39.836	0.258
Extended Cox	54.411	50.510	0.740
Extended Cox reduced	59.272	50.424	0.742
Dynamic DeepHit	47.894	37.587	0.147
Dynamic DeepHit reduced	49.950	44.367	0.019
	Sensibility	**Specificity**	
Cox	0.944	0.339	
Cox reduced	1.000	0.266	
Cox with bootstrap	0.806	0.354	
DeepHit	0.944	0.051	
DeepHit reduced	0.722	0.436	
DeepHit with bootstrap	0.931	0.245	
Extended Cox	0.278	0.749	
Extended Cox reduced	0.333	0.750	
Dynamic DeepHit	0.556	0.139	
Dynamic DeepHit reduced	0.944	0.001	

Table 2: Metrics of all the models

4. Conclusions

The majority of previous studies on the relationship between ARTs and CVD events has approached this problem with classical regression and classification methods, and only a few with survival analysis tools. In this work we compared the classic Cox Proportional Hazard model with the DeepHit, that is a recent neural network based model through their application on a real dataset, first at the baseline and then with time-dependent data. The Cox method showed that the PH assumption constrains its application, in fact, the full model at the baseline violated it. The DeepHit does not make this assumption and it is therefore more flexible, moreover it performs slightly better. The two methods show similar results in terms of the significant variables, the DeepHit model was able to capture some non-linear relationships and also to visualise them through the Shapley AE. The time-dependent data have allowed the models to reach better performances but at the same time have increased the computational cost and the results of the Dynamic DeepHit were less interpretable. The introduction of the bootstrap data showed that the complexity of neural network methods allows them, with enough data, to reach better performances. The DeepHit models showed that the time of exposure to all the ART drugs is associated with lower risk of CVD event, only a short exposure seems to increase this risk. The Cox models found that the NRTI drugs are significant and are a protective factor, note that the correlation between the ART related variables may have masked the level of significance of the other variables. Moreover we found that the patients that have started the ART before 2007 were less exposed to CVD 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 relationships between covariates and time to CVD events as they vary 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 drug on each particular CVD event.

References

[1] Clay Bavinger, Eran Bendavid, Katherine Niehaus, Richard A Olshen, Ingram Olkin, Vandana Sundaram, Nicole Wein, Mark Holodniy, Nanjiang Hou, Douglas K Owens, et al. Risk of cardiovascular disease from antiretroviral therapy for hiv: a systematic review. *PloS one*, 8(3):e59551, 2013.

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

[3] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[4] David G. Kleinbaum and Mitchel Klein. *Survival Analysis*. Springer, 3 edition, 2015.

[5] 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.

[6] 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.