



MOX-Report No. 85/2022

A neural network approach to survival analysis for modelling time to cardiovascular diseases in HIV patients with longitudinal observations

Lurani Cernuschi , A.; Masci, C.; Corso, F.; Muccini, C.; Ceccarelli, D.; San Raffaele Hospital Galli, L.; Ieva, F.; Paganoni, A.M.; Castagna, A.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it

<https://mox.polimi.it>

A neural network approach to survival analysis for modelling time to cardiovascular diseases in HIV patients with longitudinal observations

Agostino Lurani Cernuschi¹, Chiara Masci^{1,*}, Federica Corso¹, Camilla Muccini^{2,**},
Daniele Ceccarelli², Laura Galli², Francesca Ieva¹, Antonella Castagna² and Anna Maria Paganoni¹

¹MOX - Department of Mathematics, Politecnico di Milano, IT

²Department of Infectious and Tropical Diseases, IRCCS - Ospedale San Raffaele, Milano IT

email: *chiara.masci@polimi.it

email: **muccini.camilla@hsr.it

SUMMARY: At the end of 2021, 38.4 million People were Living With HIV (PLWH) worldwide. The advent of Anti Retroviral Therapy (ART) has significantly reduced the mortality and increased life expectancy of PLWH. Nowadays, the management of people with HIV on virological suppression is partly focused on the onset of comorbidities, such as the occurrence of CardioVascular Diseases (CVDs). In this study, we analyse the 15-year CVD risk in PLWH, following a survival analysis approach based on Neural Networks (NNs). We adopt a NN-based deep learning approach to flexibly model and predict the time to a CVD event, relaxing the linearity and the proportional-hazard assumptions typical of the COX model and including time-varying features. Results of this approach are compared to the ones obtained via more classical survival analysis methods, both in terms of predictive performance and interpretability, in order to explore the potential of deep learning approaches in modelling survival data with time-varying features.

KEY WORDS: CardioVascular Disease, Survival Analysis, Deep Learning, DeepHit, HIV, Neural Network, Time-Dependent Data.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

The World Health Organization (WHO) considers Human Immunodeficiency Virus (HIV) as one of the most serious public health challenges (Organization et al. 2002). AntiRetroviral Therapy (ART) has heavily decreased of 60% the expected rate of death (Pillay-van Wyk et al. 2019) of People Living With HIV (PLWH). Currently, CardioVascular Diseases (CVDs) represent one of the major causes of death among people in high-income countries (Feinstein et al. 2016) and are more common in individuals with HIV infection compared to those without (Shah et al. 2018). In fact, immune cell activation, chronic inflammation and endothelial dysfunction triggered by HIV infection itself increase the likelihood of CVDs onset (Longenecker et al. (2013); Mujawar et al. (2006)), together with a larger number of traditional risk factors, such as hypertension, dyslipidemia and smoking, more frequently observed in PLWH (Davis et al. (2021); Maggi et al. (2017)). Moreover, although more recent antiretroviral drugs seem to have a minor cardiovascular toxicity, exposure to older regimens and metabolic impairment related to current therapies may have an impact on a higher prevalence of CVDs among PLWH. For instance, Worm et al. (2010) suggested that the increment in the risk of CVDs in PLWH might be a side effect of some ART regimens. There are different classes of antiretroviral drugs: Protease Inhibitors (PIs), Nucleoside Reverse Transcriptase Inhibitors (NRTIs), Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs) and INtegrase Inhibitors (INIs), that have been introduced in 2007.

Recent studies have described potential relationship between ART and CVDs. Worm et al. (2010) found evidence from observational studies that using PIs might increase the risk of Myocardial Infarction (MI). Furthermore, Lang et al. (2010) concluded that there is no increasing risk after the exposure to any NRTIs except for abacavir and that there are no associations between MI and the exposure to NNRTIs. Machine Learning (ML)-based models are widely used in medicine to optimize diagnosis and outcome prediction, even in the field of

cardiology: indeed, ML algorithms were found to be more accurate than the American College of Cardiology/American Heart Association (ACC/AHA) cardiovascular risk assessment score, based on a parametric equation, at identifying subjects at risk of CVD (Weng et al. 2017). In HIV research, ML tools have been applied to investigate variables associated with crucial issues, including early virological suppression, neurocognitive impairment, frailty and HIV/HCV coinfection (Bisaso et al. (2018); Xu et al. (2021); Paul et al. (2020); Wei et al. (2019)). Artificial Intelligence (AI) has also proved to be effective in building public health interventions for HIV prevention, suggesting the need to combine old and new approaches to manage challenges faced in HIV (Xiang et al. (2021); Marcus et al. (2020)). However, no results are reported in the literature about the use of ML approaches in evaluating the time to the occurrence of CVDs in PLWH. Most studies in this field adopt both statistical and ML methods for classification purposes to identify the determinants of a CVD event in PLWH (Roth et al., 2021; Safo et al., 2021); other works explore the risk of CVD events in a time-to-event framework, analyzing the time to CVD events in PLWH by means of standard survival analysis (Marcus et al. (2019); D’Ascenzo et al. (2021)). While classification techniques do not use the event timing information, a survival analysis approach to this problem allows to explore the timing of CVD events and, by including time-varying covariates, to investigate the effect of the ART through time. Nonetheless, the linearity and Proportional Hazards (PH) assumptions on which classical survival models rely represent a limitation in this context. In recent years, new approaches that aim to overcome the limitations of classical survival models have been proposed. For instance, Katzman et al. (2018) developed a new survival analysis model based on NNs, called *DeepSurv*, that is able to capture non-linear relationships, still assuming proportional hazards. Kvamme et al. (2019) proposed *Cox-Time*, that fits a NN based on the Cox model with time-varying effects, where the PH assumption is no longer a restriction. Lee et al. (2018) introduced an alternative method called *DeepHit*, that uses Deep

Neural Network to estimate the distribution of survival times and allows for a time-varying relationship between covariates and risks. Lee et al. (2020) proposes an extension of DeepHit, called *Dynamic DeepHit*, that allows for the inclusion of time-varying covariates. Dynamic-DeepHit estimates the time-to-event distributions without making any assumption about the underlying stochastic model and it flexibly incorporates longitudinal data comprising various repeated measurements, in order to issue dynamically updated survival risk predictions. To the best of our knowledge, this is the only deep learning survival analysis algorithm dealing with time-varying features.

In this study, we analyze the CVD risk in PLWH followed at IRCCS¹ San Raffaele hospital in Milan, Italy, in a time to event framework. Our aim is to exploit the strength of the recently developed NN-based survival models when applied to the context of HIV research and to explore the evidences and predictive performances that these flexible methods produce, in comparison to more standard approaches. We face the problem both in a time-invariant and in a time-varying framework, comparing Cox and DeepHit approaches, including fixed and time-dependent covariates. Despite the several benefits, the main drawbacks of NN-based methods, including DeepHit, are the loss of results interpretability, the huge computational costs and the requirement of high sample sizes. In order to partially overcome these issues, we rely on recent techniques, i.e., the Permutation Feature Importance (PFI) (Breiman, 2001) and the Shapley Additive Explanation Value (Lundberg and Lee, 2017), that allow to interpret, though only qualitatively, the results and the relationship between covariates and the risk of the event.

The goal of this study is twofold: first, the comparison between new NN-based approaches to survival analysis and the well-known Cox model, with a focus on weaknesses and potentialities of these methods in terms of robustness in prediction, interpretability of results

¹Istituti di Ricovero e Cura a Carattere Scientifico (IRCCS).

and computational costs; second, the evaluation of NN-based approaches to survival analysis when applied for a clinical purpose. To the best of our knowledge, this is the first time that DeepHit and Dynamic DeepHit are applied to longitudinal time-to-event health data.

The article is organized as follows: in Section 2 we give a comprehensive description of the dataset; Section 3 recalls the mathematical formulation of Cox PH models and DeepHit, including their extended version for longitudinal data, and describes the goodness-of-fit metrics. Section 4 shows the results of Cox PH models and DeepHit applied to HIV patients data, both in a time-invariant and time-dependent framework, and compare them in terms of performance, interpretability and computational cost. Since the San Raffaele's dataset contains a very low percentage of CVD events, we tackle the problem of imbalanced data by proposing, as an additional analysis, a stratified bootstrapped version of our dataset to observe models performance on this augmented dataset. Concluding remarks are discussed in Section 5, in which we describe potentialities and limitations raised in this work.

The analysis is conducted in R (R Core Team, 2021) and Python (Van Rossum and Drake Jr, 1995) softwares.

2. IRCCS San Raffaele HIV patients dataset

CSL HIV is a cohort of adult PLWH followed at San Raffaele Hospital in Milan, Italy (the CSL HIV Cohort Study) that contains data about 4512 patients included in the analysis, collected between 1998 and 2021².

The target variable, for each patient, is the time from the ART beginning (baseline) until the occurrence of a CVD event, if any. Only cardiovascular diseases that occurred within 15 years are considered for the analysis, whereas patients with a CVD event occurred after 15 years are censored (Kleinbaum and Klein, 2010).

²We do not consider data before 1998 since at that time the ART was not a combination of the four classes of drugs considered in this dataset yet.

For each patient, the clinical history, i.e. the collection of information measured at each visit from the ART beginning, is registered, including demographic variables (e.g., sex, race, age, etc.), clinical parameters (e.g., viremia, cholesterol, etc.) and cumulative time exposure to ART drugs.

In accordance with the clinician's experience and to overcome some limitations induced by the proportion of missing data, we select 21 variables of interest (7 categorical and 14 numeric). Specifically, 3 of them are time-invariant and the remaining ones are longitudinal. Among the 6 binary variables, the 4 longitudinal ones are defined as step-functions: the diagnosis of tumor, diabetes, acquired immune deficiency syndrome (AIDS) and hypertension. The complete list and the description of the variables are reported in Table 1. Summary statistics of categorical and numerical covariates are reported in Table 2.

3. Methods

3.1 Basics of survival analysis

In survival analysis, for each observation $i = 1, \dots, N$, the target variable is defined as the couple of the survival time $T_i = \min(T_i^*, C_i)$ and the censoring indicator $\delta_i = \mathbb{1}(T_i^* \leq C_i)$, where C_i is the censoring time and T_i^* is the CVD event time, if any. δ_i is the indicator function that indicates whether the event occurred ($\delta_i = 1$) or not ($\delta_i = 0$) for the individual i . Censoring is assumed independent of survival time. The survival function $S(t_i) = \mathbb{P}(T > t_i) = 1 - \mathbb{P}(T \leq t_i) = 1 - F(t_i)$ represents the probability of survival until time t_i . From this definition, we can extract the hazard function:

$$h(t_i) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t_i \leq T \leq t_i + \Delta t | T \geq t_i)}{\Delta t} \quad (1)$$

that describes the instantaneous risk of failure. The survival function $S(t)$ of a group of patients can be estimated through the Kaplan-Meier estimator (KM), that represents the probability of surviving in a given length of time while considering time in many small

intervals. In case of two or more groups, the Log-Rank Ratio test can be used to test statistical differences across the estimated KM curves. The KM estimator has also been extended to the case of time-dependent variables.

3.2 The Cox model

The Cox PH model is one of the most used regression models in survival analysis. It is a semi-parametric model that studies the effect of a set of covariates \mathbf{x}_i , relative to an individual i on the instantaneous risk $h_i(t)$ of the event to occur. The Cox PH model expresses the hazard function for an individual i as:

$$h_i(t|\mathbf{x}_i) = h_0(t)e^{\mathbf{x}_i^T\boldsymbol{\beta}} \quad (2)$$

where $h_0(t)$ is the unspecified baseline hazard function and $\boldsymbol{\beta}$ is the unknown vector of regression coefficients. The model parameters are estimated via maximization of the partial likelihood (Kleinbaum and Klein, 2010).

For a set of time-dependent covariates, the extended Cox model (Therneau and Grambsch, 2000) assumes the hazard function at time t to depend only on the covariates measured at time t . Given the covariates $\mathbf{x}_i(t)$ for the i^{th} individual, the hazard function is modelled as:

$$h_i(t|\mathbf{x}_i(t)) = h_0(t)e^{\left[\sum_{p=1}^{P_{fix}} x_{ip}\beta_p + \sum_{p=P_{fix}+1}^{P_{fix}+P_{td}} x_{ip}(t)\beta_p\right]} \quad (3)$$

where \mathbf{x}_i are the covariates of the i^{th} individual, both time-dependent and invariant; P_{fix} and P_{td} are the numbers of time-invariant and time-varying covariates, respectively; $\boldsymbol{\beta}$ is the vector of coefficients. As for the standard Cox model, the coefficients of time-dependent covariates are estimated using a maximum likelihood approach, that provides one coefficient for each longitudinal covariate. This coefficient represents the total effect of the corresponding time-dependent variable, considering all the times at which the variable was measured. In

this setting, the PH assumption is no longer satisfied, being the hazard ratios non-constant over time.

3.3 DeepHit

DeepHit is a NN-based method for survival analysis (Lee et al., 2018). Its architecture includes a first network that captures the features common to all individuals experiencing an event and K subsequent parallel cause-specific networks. Each of the K cause-specific networks focuses on a particular competing risk. The output layer is a Softmax function (Bridle, 1990) that returns the joint distribution of the probability of event for each class of competing risk. The network is able to capture the relationships between covariates and risks, which can be non-linear and non-proportional over time.

In case of time-invariant variables, the architecture of DeepHit is basically composed by Feed Forward Neural-Networks (FFNNs)(Bebis and Georgiopoulos, 1994). Each FFNN has different numbers of hidden layers with a numerous set of parameters to be found by minimizing a loss function, that, for this model, is the sum of the negative *log* likelihood L_1 and the ranking loss function L_2 . L_1 handles censored data and, by minimizing it, the algorithm learns to identify the differences in the features of censored and non-censored data. Given the (cause-specific) Cumulative Incidence Function (CIF) for the event $k = k^*$, defined as $F_{k^*}(t^*|\mathbf{x}^*) = \mathbb{P}(T \leq t^*, k = k^*|\mathbf{x}^*)$, which estimates the probability that the event k^* occurs on or before time t^* conditional to the covariates \mathbf{x}^* , then the estimated output layer for the event k^* is $y_{k^*,t^*} = \widehat{P}(t^*, k^*|\mathbf{x}^*)$. By summing up $y_{k^*,\tilde{t}}, \forall \tilde{t} = 1, \dots, t^*$, the estimated CIF $\widehat{F}_{k^*}(t^*|\mathbf{x}^*)$ is obtained. The *log* likelihood of the joint distribution of the first hitting time and corresponding event L_1 is defined as :

$$L_1 = - \sum_{i=1}^N [\mathbf{I}_{\{k^i \neq \emptyset\}} \log(y_{k^i, t^i}^i) + \mathbf{I}_{\{k^i = \emptyset\}} \log(1 - \sum_{k=1}^K \widehat{F}_k(t^i|\mathbf{x}_i))]. \quad (4)$$

The first term of the Eq. (4) regards censored observations and the second term the uncen-

sored ones. Since a survival model is assumed to perform well if, comparing two individuals, it predicts the one with lower hitting time with a higher probability of the event, the ranking loss function L_2 is introduced in order to teach the model to predict the correct ordering of patients, by following the concordance idea. For instance, a patient who dies at time t should have a higher risk at time t than a patient who survived longer than t . $A_{k,i,j} = \mathbb{1}(k^i = k, k^j = k, t^i < t^j)$ is the indicator function for each pair (i, j) of individuals who experience risk k at different times, and whose risks for event k can therefore be directly compared. The ranking loss function L_2 is:

$$L_2 = \sum_{k=1}^K \alpha_k \sum_{i \neq j} (A_{k,i,j} \eta(\widehat{F}_k(t^i | \mathbf{x}_i), \widehat{F}_k(t^j | \mathbf{x}_j))) \quad (5)$$

where the parameters α_k are chosen to trade off ranking losses of the k^{th} competing event, and $\eta(x, y) = \exp(\frac{x-y}{\sigma})$ is a convex loss function.

For each network, the selection of hyper-parameters (e.g., number of hidden layers and neurons for each layer) is obtained via estimation of the optimal minimum of the loss function, controlling for overfitting.

Overfitting is an important issue, especially in strongly unbalanced cases. Two possible solutions to prevent it are the use of dropout (i.e., set a random percentage of parameters to zero) and the weight regularization (i.e., constrain parameters weight to a maximum value).

By introducing a new network and a further loss function component, the DeepHit algorithm can be extended to handle time-dependent variables (Dynamic DeepHit, Lee et al. (2020)). The Dynamic DeepHit has the same layers structure of FFNNs, but each neuron is auto-connected to itself, allowing the network to process old data at each instant. A Long Short-Term Memory (LSTM) network solves the problem of vanishing gradient, i.e., when the propagation through the Recurrent NN (RNN) leads weights to zero, making the model non-trainable.

The Dynamic DeepHit uses a new loss function L_3 defined as:

$$L_3 = \beta \cdot \sum_{i=1}^N \sum_{j=0}^{J^i-1} \sum_{d \in \mathcal{I}} (1 - m_{j+i,d}^i) \cdot \mathcal{C}(x_{j+1,d}^i, y_{j,d}^i) \quad (6)$$

where J^i is the number of time stamps for the i -th individual, $m_{j+i,d}^i$ shows if the covariate d at time t_{j+1} for the i^{th} patient is missing, β is a hyper-parameter that regulates the importance of the time variability of data during the training, $\mathcal{C}(a, b) = |a - b|^2$ for continuous variables and $\mathcal{C}(a, b) = -a \log(b) - (1 - a) \log(1 - b)$ for binary variables. \mathcal{I} is the set of time-varying covariates on which the network is regularized. This makes predictions on the one step-ahead covariate $x_{j+1,d}$ to regularize the shared sub-network so that the hidden representations preserve information for the step-ahead predictions. An important advantage of dynamic DeepHit is the ability of making the prediction by considering all the measures until a specific time point and not only by observing the covariate value at that time point as in classical survival methods.

3.3.1 Deephit interpretation. Since Deephit does not provide Hazard Ratios (HRs) and p-values as the Cox model does, alternative methods that serve to extract information from the results of NN-based methods have been recently proposed in the literature. Breiman (2001) proposed the Permutation Feature Importance (PFI) and Lundberg and Lee (2017) suggested the Shapley Values (SVs). PFI computes a ranking of the covariates importance, giving insights about how much each covariate contributes to the model prediction. In this case, a positive contribution is measured as an increment in the C-index (see Section 3.4). Although this method helps in the covariates selection, it does not reveal the type of association between each covariate and the target variable. Shapley values give insights about the type of effect (i.e., increase or decrease the risk) that each covariate gives to the prediction of the target variable (i.e. risk of the event). For instance, the marginal contribution of a covariate is measured as the difference between the classic prediction (i.e., the one obtained by considering the complete set of covariates) and the one obtained by randomly shuffling

the values of the covariate. The SV is then calculated as the expected difference among all the differences obtained by replicating the process many times. Thus, a positive SV for a covariate indicates that an increment in the values of the covariate contributes to an increment in the predicted risk of the event. Moreover, SVs can be used to evaluate the effect of a single covariate as well as the one of the interaction of two covariates, evaluating their joint effect.

3.4 Goodness-of-fit metrics

The Concordance index (C-index) (Harrell et al., 1982) is one of the most used metrics to evaluate survival models. It evaluates the capability of the model to predict the correct ordering of events relative to individuals, by computing the percentage of individuals couples that are predicted concordant to the true ordering of survival times. A pair of observations are defined concordant if the individual with lower survival time is predicted with higher risk of event than the other. The formula of the C-index is the following:

$$\text{C-index} = \frac{\sum_i \sum_j \mathbb{1}(T_i < T_j) \mathbb{1}(\delta_i = 1) \mathbb{1}(\text{risk}_i > \text{risk}_j)}{\sum_i \sum_j \mathbb{1}(T_i < T_j) \mathbb{1}(\delta_i = 1)} \quad (7)$$

for $i, j = 1, \dots, N$, where T_i is the observed survival time and risk_i is the predicted instantaneous risk for the individual i . It takes values between 0 and 1: a 0 value is relative to an inverse predictor, a 0.5 value to a random predictor and a 1 value to an ideal predictor. Its main limitation is that it does not take into account whether two individuals have a similar or very dissimilar time-to-event. It is easier to predict the correct ordering of two patients with very far times-to-event than the one of two patients with very close times-to-event. A mistake in the prediction for the first couple should be considered more serious than a mistake in the second one, but the C-index does not take into account this aspect. Given this limitation, to complement the C-index, we rely on other metrics to properly evaluate the performance of the models. We define the standard Mean Square Error (MSE) estimated all over the uncensored

individuals. The estimated survival time is the time in which the estimated survival curve meets a threshold, that is chosen to minimize the MSE on the training set. This threshold is then used to compute the MSE on the test set. Lastly, by approaching the problem as a classification problem, a further evaluation metric regards the computation of the Receiving Operating Characteristic (ROC) curve. The classification threshold \tilde{p} is chosen to maximize the geometric mean between the sensitivity and the specificity. This metric is particularly suited for imbalanced classification problems (Barandela et al., 2003). The threshold \tilde{p} is estimated on the training set and then used to evaluate the model on the test set. In this binary classification setting, the performance evaluation is supported by the the Area Under the ROC (AUROC) and by indexes of accuracy, sensitivity and specificity.

4. Results

Overall, 4512 PLWH are evaluated: 80% are males, the median age is 37.5 years at the baseline (IQR 31.37-44.37). During a median follow-up of 16.5 (IQR 10.7-22.8) years, 90 (2%) PLWH experienced a CVD event in 15 years. The mean of the time-to-CVD event is 7.19 years (sd = 4.09) and the average number of visits for each patient is 32.01 (sd = 28.19). The distributions of the time-to-CVD event and of the number of visits for each patients are reported in Figure S1 and S2 in Supplementary materials, respectively.

In Subsection 4.1, we present results of models fitted at the baseline, i.e. considering patients clinical and personal information measured at the beginning of the ART, while in Subsection 4.2, we present results of models that consider the entire patients clinical history, by means of time-dependent covariates (e.g., time of exposure to drugs, time of a disease diagnosis,..). Results of Cox models and DeepHit are compared in terms of predictive performance and interpretability.

4.1 Models at baseline

We approach the time-to-CVD events study by considering patients information at the beginning of the ART, i.e. at the baseline³. For each covariate, a univariate explorative analysis is conducted through KM estimator. Numerical variables have been discretised standing on clinical cut-offs or median values. At the baseline, the significant variables, according to the log-rank test, result to be age, diabetes, hypertension, hepatitis C (HCV), CD4 T cells, hemoglobin, creatinine, race, platelets and AST. In particular, it emerges that, at the baseline, the diagnosis of Hepatitis C, hypertension and diabetes, high values of creatinine and ASpartate aminoTransferase ($> 35 \frac{UI}{L}$), low values of hemoglobin ($< 12 \frac{g}{dL}$) and white race are risk factors, while being young, high values of CD4 T cells ($> 200 \frac{cell}{mcL}$) and platelets ($> 200000/mm^3$) are protective factors. Figure S3 in Supplementary materials reports the KM curves for these significant features.

4.1.1 Cox PH model. The multivariate Cox PH model is fitted using the 17 covariates measured at the baseline (listed in Table 1, excluding the ART-related covariates) and its results are reported in left panel of Figure 1. Among the 17 covariates, the significant ones result to be HCV, hypertension, age and creatinine. Given the low percentage of events and the high number of covariates, we proceed with a stepwise procedure to reduce the set of covariates and build a more parsimonious and robust model. The reduced model retains 5 covariates and its output is shown in right panel of Figure 1. The five significant covariates are consistent with the four ones identified in the full Cox model plus CD4. In particular, being an older patient at the beginning of the ART (HR=1.05), the presence of hepatitis C (HR=2.57), the diagnosis of hypertension (HR=3.26) and an increment in creatinine level (HR=1.61) result to be risk factors, while having high levels of CD4 T cells is a protective factor (HR=0.9899).

³Since the exposure time to ART drugs at the baseline is zero, we do not include any ART-related covariate.

4.1.2 *DeepHit*. After a fine tuning process to select the NN parameters, we run the DeepHit algorithm at baseline with a dropout percentage of 50% and a maximum parameter weight of 0.001. Since DeepHit does not directly return information about the covariates significance, its results are interpreted by means of PFI and SVs. The five most important variables result to be age, platelets, CD4, hypertension and cholesterol (top panels of Figure 2). To investigate whether these are risk or protective factors, we visualize the relationship between covariates values and risk of CVD, together with a covariates importance ranking, by means of SVs. The most important covariates and their qualitative relationships with the risk of CVD event are reported in the top right panel of Figure 2. The most important variables are consistent with the ones identified by the Cox model at baseline, including age, HCV, hypertension, triglycerides, CD4, viremia and platelets. According to these results, a reduced DeepHit model including only these 7 covariates is built. Results of SVs and PFI applied to the reduced DeepHit are reported in bottom panels of Figure 2. The covariates ordering follows the feature importance that is evaluated as the mean of the SVs for each covariate. Right bottom panel of Figure 2 shows that patients with high values of age, triglycerides and viremia are more likely to experience a CVD event. Having a diagnosis of hypertension or HCV increases the risk of event, while greater values of CD4 and platelets decreases it. Figure S4 in Supplementary materials shows the SVs for the joint effect of age and CD4. The age of patients is represented through a scaled color from older patients in red to the younger in blue. Old patients, with a low level of CD4 represent the group at highest risk (positive SVs), while, as we expected, patients with high levels of CD4 are less likely to experience a CVD event, independently of the age.

4.1.3 *Goodness-of-fit and predictive power*. First part of Table 3 reports the C-index relative to the four models fitted at the baseline, computed on the training (80% of the data) and on the test set (remaining 20% of the data), together with MSE on the test set, accuracy

and sensitivity. In terms of C-index, the two approaches reach similar results and, for both of them, reduced models perform better than full ones. The C-index on the test set of the full Cox model and DeepHit are very similar (0.6432 and 0.6443 respectively), but the reduced Cox model results to perform the best in terms of C-index, MSE and sensitivity. The full DeepHit AUROC at baseline (that coincides with the reduced DeepHit one) is slightly better than the full Cox PH model one (0.72 with $\tilde{p} = 0.033$ and 0.66 with $\tilde{p} = 0.0061$ respectively). The full DeepHit reaches the highest accuracy (0.7320), that decreases reducing the set of covariates (0.6246), while, the sensitivity increases moving from the full to the reduced set of covariates (0.4444 and 0.6667, respectively). The full and reduced Cox models reach a similar accuracy (0.6202 and 0.6224, respectively), but the reduced model performs better than the full one in terms of sensitivity (0.7222 and 0.6667, respectively). In terms of MSE, the Cox PH models perform better than DeepHit ones.

The estimated survival curves of the four models, where the survival probability is intended as the probability of not having a CVD event, computed on the training set and stratified by patients with and without the event, are reported in Figure 3. Both Cox and DeepHit estimate a lower median survival curve for patients who actually experience the event with respect to the ones who do not. Nonetheless, the delta between the median survival curves, measured at 15 years, between patients who do experience the event and patients who do not is much more pronounced in the DeepHit prediction than in the Cox one. From this perspective, DeepHit performs better in recognizing distinguished survival trends in the two types of HIV patients.

4.1.4 Bootstrapped dataset. Given the low number of observations with event within the test set (18 patients), we conduct a bootstrap analysis by taking the 60% of the data as training set and bootstrapping it as 7-times bigger than the original one. In particular, we sample data with replacement from the original dataset preserving the underlying distribu-

tion and the proportion of censored data. A jittering noise is also added to avoid overfitting and to increase the variability within the new dataset. After a tuning process, we sample a noise with a variance equal to the 25% of the variance of each feature. In this way, the test set is made by the remaining 40% of the data (36 patients experiencing the event) and the evaluation of the C-index may gain in reliability. We run Cox and DeepHit models on the bootstrapped dataset considering the reduced set of covariates selected by the reduced Cox model at baseline (i.e., HCV, hypertension, CD4, age, creatinine). Models performances are reported in the second part of Table 3. On the test set, Cox and DeepHit methods fitted on the bootstrapped dataset reach very high values of C-index (0.7853 for Cox and 0.7882 for DeepHit) and of the AUROC (0.76 with $\tilde{p} = 0.036$ for Cox and 0.75 with $\tilde{p} = 0.054$ for DeepHit). These results suggest that the strong imbalance in the target variable and the low sample size of the data might have heavily affected the previous performances of DeepHit.

4.2 Models with time-dependent covariates

We now extend the models to include time-varying covariates, which regard the clinical history of each patient, administered over time, and the cumulative exposure time to ART drugs, which enables us to study how different combinations of drugs are associated to the risk of CVD events. We measure an average number of 32 visits for each patient meaning that a high computational cost is needed to fit models with time-dependent data. The univariate analysis by means of KM estimator for time-varying variables reveal results similar to the ones at the baseline. Moreover, we add now the information of the ART-related variables. Bavinger et al. (2013) argued about the impact of the timing of exposure to ART inhibitors by sustaining that this exposure has different impacts on the CVD risk standing on whether it is a recent exposure or a later exposure, i.e. shorter or longer than 6 months. Therefore, we choose 6 months for the cut-off for the KM estimator referred to the exposure time of

inhibitors but none of the resulting curves suggest a difference in the risk of CVD event between the two exposure times.

4.2.1 Time-dependent Cox model. As we did for models at the baseline, we first fit a full model (including the 21 covariates listed in Table 1) and then we reduce it by means of a stepwise procedure, retaining the most significant covariates. Results of the full and reduced Cox models are reported in left and right panels of Figure 4, respectively. In the full model, the most significant variables result to be age, hypertension, HCV, creatinine (that were also significant at the baseline), together with cholesterol and triglycerides. The reduced model, in addition to these six covariates, identifies as significant also the time of exposure to NRTIs, that results to be a protective factor with HR equal to 0.94.

4.2.2 Dynamic DeepHit. Due to prohibitive computational time and cost needed to compute SVs in case of longitudinal data, dynamic DeepHit results are interpreted by means of PFI, using the C-index on the training set. Results of full and reduced Dynamic DeepHit are reported in left and right panels of Figure 5, respectively⁴. The reduced model is built by including the most relevant features from both the full time-dependent Cox and Dynamic DeepHit models. Platelets, NRTIs exposure time, hypertension, Hepatitis C and age result to be the most important predictors.

4.2.3 Goodness-of-fit and predictive power. Third part of Table 3 reports the performance of models with time-varying covariates. Overall, the performance of these models is higher than the one of models at the baseline, as expected. The C-index measured on the test set suggests that the reduced Cox model (C-index 0.7059) performs better than the full one, that eventually suffers from overfitting (C-index 0.6854). The full Dynamic DeepHit

⁴For the Dynamic DeepHit, the fine tuning process for the NN parameters selected 20% as the best value for the dropout percentage and 0.00001 for the weight regularization.

model reaches the best performance (C-index 0.7104) on the test set. The highest AUROC is reached by the full dynamic DeepHit (0.83 with $\tilde{p} = 0.027$) followed by its reduced version (0.78 with $\tilde{p} = 0.011$), whereas AUROCs of full and reduced Cox models remain similar to the ones at the baseline (0.68 with $\tilde{p} = 0.018$ and 0.71 with $\tilde{p} = 0.013$, respectively). In terms of accuracy and sensitivity, DeepHit models reach higher results. The full Dynamic DeepHit model produces, on the test set, a sensitivity that is equal to the one of the full Cox model (0.7778), but a greater accuracy (0.7110 vs 0.4983). The reduced Cox model reaches the best performance in terms of sensitivity (0.9444) but, as drawback, it predicts with low accuracy (0.3699).

On the contrary, the reduced Dynamic DeepHit predicts with the highest accuracy (0.7796) but with low sensitivity (0.5556), resulting to be less sensitive in identifying patients at risk. MSEs are generally lower than the ones estimated at the baseline. The lowest MSE is achieved by the reduced time-dependent Cox model (17.6951).

For time-dependent models, the distributions of the estimated survival curves, stratified by patients with and without the event, are reported in Figure 6. All models estimate a lower median survival curve for those patients who actually experienced the event, with respect to the others. In the time-dependent setting, a relevant improvement of Dynamic Deephit with respect to Cox model regards the possibility to consider the entire clinical history of a patient when making a prediction. Time-dependent Cox models make predictions at a specific time instant by looking only at the covariates' values in this specific time instant, neglecting all previous values. On the contrary, Dynamic DeepHit uses all records until the specific time instant, leading to a more informed predictor. This leads to a better fit, as can be observed in bottom panels of Figure 6.

5. Discussion

In this study, we explore the potential of innovative NN-based survival methods, DeepHit and Dynamic DeepHit, when applied in HIV research for predicting the 15-year CVD risk in PLWH. We compare their results with the ones of standard Cox models, both in terms of interpretability and predictive performance. We face two different settings: a time-invariant one, in which we consider time-invariant features, and a more complex time-varying one, in which we add longitudinal features.

When considering the only time-invariant features measured at the beginning of the ART, Cox and DeepHit identify very similar sets of protective and risk factors for the risk of CVD, that comprehend age, level of CD4 T cells, diagnosis of hypertension and HCV. The predictive performances of the two methods are also very close, with DeepHit performing slightly better in terms of AUROC and related indices, but worse in terms of MSE. When moving to the time-dependent setting, we particularly focus on the exposure time to antiretroviral drugs, since the relationship between ART and CVD event is receiving particular attention in the literature (Worm et al., 2010). To the best of our knowledge, this aspect has been investigated through classical regression and classification methods, while applications of time-dependent NNs for survival analysis are still rare in the literature. Both Cox and DeepHit identify again age, level of CD4 T cells, diagnosis of hypertension and hepatitis C as predictive factors, together with triglycerides and exposure time to NRTI, that, in particular, results to be a protective factor. Cox identifies exposure time to NRTI as the only relevant ART related variable, while Dynamic DeepHit identifies, in addition to it, also exposure time to PI. In terms of predictive performances, in the time-dependent setting, DeepHit confirms its superiority in terms of AUROC and related indices but also its weakness in terms of MSE. Nonetheless, DeepHit shows a very good ability in estimating distinguished survival curves for the populations of PLWH with and without the event. From a clinical point of view, no

evidence of an impact of exposure time to ART on the risk of CVD is found. Considering that CVDs are more frequently observed among PLWH than in general population, this study confirms that the variables involved in the prediction of cardiovascular events are multiple, both related to HIV infection and traditional risk factors.

Overall, DeepHit and Dynamic DeepHit present both advantages and disadvantages when applied to this context. Despite NN-based methods have the advantage of not assuming proportional hazards and of capturing non-linear relationships between covariates and target variable, the difficulty in discovering the nature of these potentially varying and non-linear relationships represents a big limitation from the interpretability point of view. Moreover, a correct implementation of the algorithm requires to tune the hyperparameters and to control for the overfitting, that is very time consuming. Nonetheless, we find that generally the Cox model suffers more in robustness, when a big set of covariates is considered, while techniques such dropout and weight regularization, used in the NN, works better against overfitting. Furthermore, the dynamic DeepHit, unlike Cox method, predicts a survival time point by including all the clinical measurements recorded for a patient until that point, providing a more learned predictor. Based on these findings, ML approaches may have a role in tailoring HIV management and offering personalized care.

The main limitation of this work regards the problems related to the data unbalance. The proportion of patients experiencing a CVD event is extremely low and this induces a further challenge for the application of the algorithms, especially the NN-based ones. The Bootstrap analysis, although only partially, gives us some insights about the improvements that NN-based methods might show in presence of more numerous and balanced data.

This work represents a further step in the rising field of ML survival methods in health research. Further works should to be designed using bigger datasets in order to develop more sophisticated tools able to enhance disease risk prediction in clinical practice.

SUPPLEMENTARY MATERIALS

Web Appendix, referenced in Section 4, is available with this paper at the Biometrics website on Wiley Online Library.

REFERENCES

- Barandela, R., Sánchez, J. S., García, V., and Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition* **36**, 849–851.
- Bavinger, C., Bendavid, E., Niehaus, K., Olshen, R. A., Olkin, I., Sundaram, V., Wein, N., Holodniy, M., Hou, N., Owens, D. K., et al. (2013). Risk of cardiovascular disease from antiretroviral therapy for hiv: a systematic review. *PloS one* **8**, e59551.
- Bebis, G. and Georgiopoulos, M. (1994). Feed-forward neural networks. *IEEE Potentials* **13**, 27–31.
- Bisaso, K. R., Karungi, S. A., Kiragga, A., Mukonzo, J. K., and Castelnuovo, B. (2018). A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating hiv patients. *BMC medical informatics and decision making* **18**, 1–10.
- Breiman, L. (2001). Random forests. *Machine learning* **45**, 5–32.
- Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202.
- D’Ascenzo, F., De Filippo, O., Gallone, G., Mittone, G., Deriu, M. A., Iannaccone, M., Ariza-Solé, A., Liebetrau, C., Manzano-Fernández, S., Quadri, G., et al. (2021). Machine learning-based prediction of adverse events following an acute coronary syndrome (praise): a modelling study of pooled datasets. *The Lancet* **397**, 199–207.

- Davis, K., Perez-Guzman, P., Hoyer, A., Brinks, R., Gregg, E., Althoff, K. N., Justice, A. C., Reiss, P., Gregson, S., and Smit, M. (2021). Association between hiv infection and hypertension: a global systematic review and meta-analysis of cross-sectional studies. *BMC medicine* **19**, 1–16.
- Feinstein, M. J., Bahiru, E., Achenbach, C., Longenecker, C. T., Hsue, P., So-Armah, K., Freiberg, M. S., and Lloyd-Jones, D. M. (2016). Patterns of cardiovascular mortality for hiv-infected adults in the united states: 1999 to 2013. *The American journal of cardiology* **117**, 214–220.
- Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama* **247**, 2543–2546.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53**, 457–481.
- Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology* **18**, 1–12.
- Kleinbaum, D. G. and Klein, M. (2010). *Survival analysis*, volume 3. Springer.
- Kvamme, H., Borgan, Ø., and Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *arXiv preprint arXiv:1907.00825* .
- Lang, S., Mary-Krause, M., Cotte, L., Gilquin, J., Partisani, M., Simon, A., Boccara, F., Costagliola, D., of the French Hospital Database on HIV, C. E. G., et al. (2010). 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**, 1228–1238.
- Lee, C., Yoon, J., and van der Schaar, M. (2020). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data.

IEEE Transactions on Biomedical Engineering (TBME) .

Lee, C., Zame, W. R., Yoon, J., and van der Schaar, M. (2018). Deephit: A deep learning approach to survival analysis with competing risks. *Conference on Artificial Intelligence (AAAI)* .

Longenecker, C. T., Funderburg, N. T., Jiang, Y., Debanne, S., Storer, N., Labbato, D. E., Lederman, M. M., and Mccomsey, G. A. (2013). Markers of inflammation and cd8 t-cell activation, but not monocyte activation, are associated with subclinical carotid artery disease in hiv-infected individuals. *HIV medicine* **14**, 385–390.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30**,

Maggi, P., Di Biagio, A., Rusconi, S., Cicalini, S., D’Abbraccio, M., d’Ettorre, G., Martinelli, C., Nunnari, G., Sighinolfi, L., Spagnuolo, V., et al. (2017). Cardiovascular risk and dyslipidemia among persons living with hiv: a review. *BMC Infectious Diseases* **17**, 1–17.

Mantel, N. et al. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* **50**, 163–170.

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

Marcus, J. L., Sewell, W. C., Balzer, L. B., and Krakower, D. S. (2020). Artificial intelligence and machine learning for hiv prevention: emerging approaches to ending the epidemic. *Current HIV/AIDS Reports* **17**, 171–179.

Mujawar, Z., Rose, H., Morrow, M. P., Pushkarsky, T., Dubrovsky, L., Mukhamedova, N., Fu, Y., Dart, A., Orenstein, J. M., Bobryshev, Y. V., et al. (2006). Human immunodeficiency virus impairs reverse cholesterol transport from macrophages. *PLoS biology* **4**, e365.

- Organization, W. H. et al. (2002). The work of who on hiv. Technical report, World Health Organization.
- Paul, R. H., Cho, K. S., Luckett, P., Strain, J. F., Belden, A. C., Bolzenius, J. D., Navid, J., Garcia-Egan, P. M., Cooley, S. A., Wisch, J. K., et al. (2020). Machine learning analysis reveals novel neuroimaging and clinical signatures of frailty in hiv. *Journal of acquired immune deficiency syndromes (1999)* **84**, 414.
- Pillay-van Wyk, V., Msemburi, W., Dorrington, R., Laubscher, R., Groenewald, P., and Bradshaw, D. (2019). Hiv/aids mortality trends pre and post art for 1997-2012 in south africa—have we turned the tide? *South African Medical Journal* **109**, 41–44.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Roth, J. A., Radevski, G., Marzolini, C., Rauch, A., Günthard, H. F., Kouyos, R. D., Fux, C. A., Scherrer, A. U., Calmy, A., Cavassini, M., et al. (2021). 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**, 1198–1208.
- Safo, S., Haine, L., Baker, J., Reilly, C., Duprez, D., Neaton, J., Wang, J., Jain, M. K., Pinto, A. A., Staub, T., et al. (2021). 89976 assessing protein biomarkers role in cvd risk prediction in persons living with hiv (pwh). *Journal of Clinical and Translational Science* **5**, 47–48.
- Shah, A. S., Stelzle, D., Lee, K. K., Beck, E. J., Alam, S., Clifford, S., Longenecker, C. T., Strachan, F., Bagchi, S., Whiteley, W., et al. (2018). Global burden of atherosclerotic cardiovascular disease in people living with hiv: systematic review and meta-analysis. *Circulation* **138**, 1100–1112.
- Snapinn, S. M., Jiang, Q., and Iglewicz, B. (2005). Illustrating the impact of a time-varying

- covariate with an extended kaplan-meier estimator. *The American Statistician* **59**, 301–307.
- Therneau, T. M. and Grambsch, P. M. (2000). The cox model. In *Modeling survival data: extending the Cox model*, pages 39–77. Springer.
- Van Rossum, G. and Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.
- Wei, Y., Li, W., Du, T., Hong, Z., and Lin, J. (2019). Targeting hiv/hcv coinfection using a machine learning-based multiple quantitative structure-activity relationships (multiple qsar) method. *International journal of molecular sciences* **20**, 3572.
- Weng, S. F., Reys, J., Kai, J., Garibaldi, J. M., and Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one* **12**, e0174944.
- Worm, S. W., Sabin, C., Weber, R., Reiss, P., El-Sadr, W., Dabis, F., De Wit, S., Law, M., Monforte, A. D., Friis-Møller, N., et al. (2010). 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**, 318–330.
- Xiang, Y., Du, J., Fujimoto, K., Li, F., Schneider, J., and Tao, C. (2021). Application of artificial intelligence and machine learning for hiv prevention interventions. *The Lancet HIV*.
- Xu, Y., Lin, Y., Bell, R. P., Towe, S. L., Pearson, J. M., Nadeem, T., Chan, C., and Meade, C. S. (2021). Machine learning prediction of neurocognitive impairment among people with hiv using clinical and multimodal magnetic resonance imaging data. *Journal of neurovirology* **27**, 1–11.

TABLES

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

FIGURES

[Figure 1 about here.]

[Figure 2 about here.]

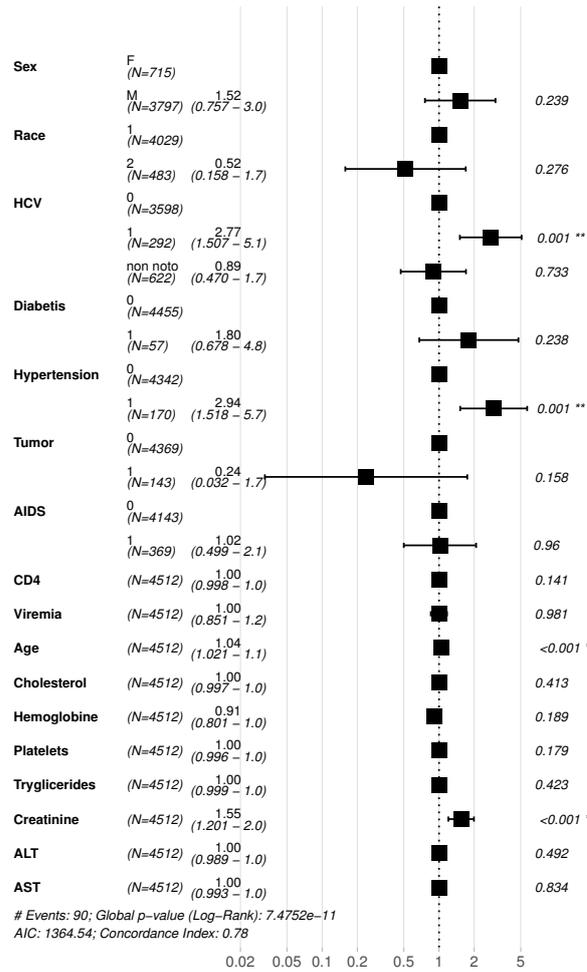
[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

Full Cox model at the baseline
Hazard ratio



Reduced Cox model at the baseline
Hazard ratio

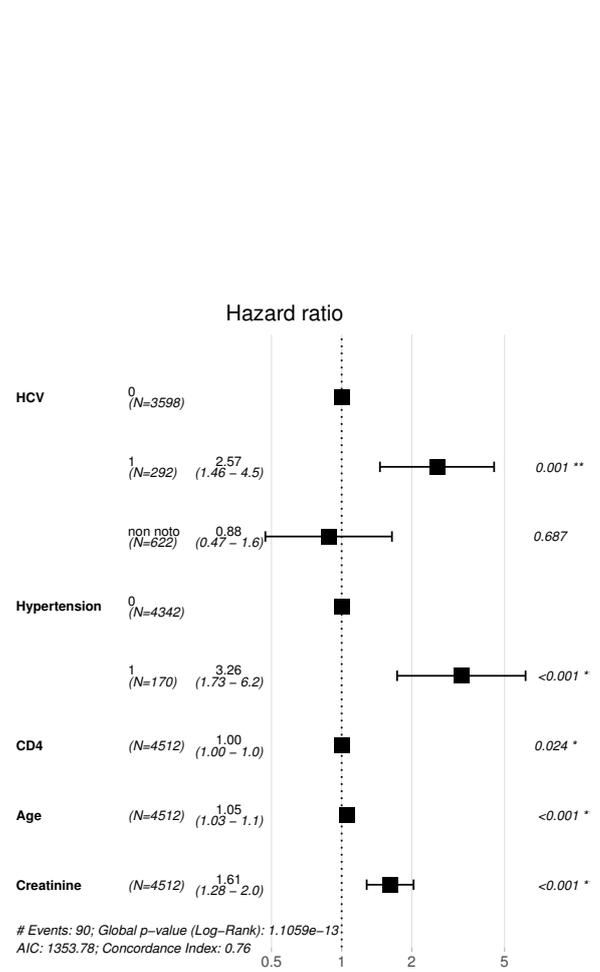


Figure 1. Results of the Cox PH model fitted with the full (left panel) and the reduced (right panel) set of covariates measured at the baseline.

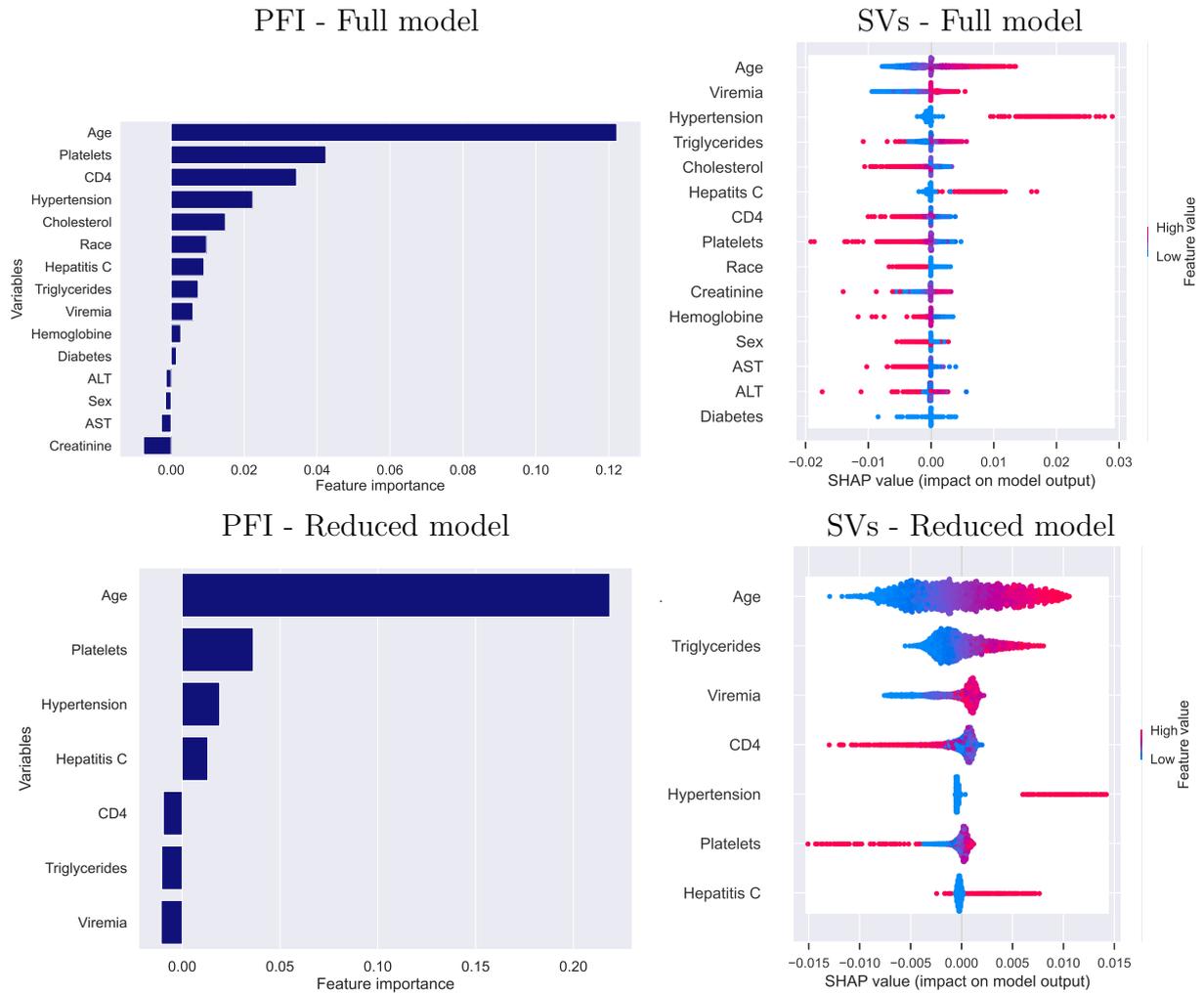


Figure 2. Features importance of the DeepHit full (top panels) and reduced (bottom panels) models at the baseline obtained with PFI (left panels) and SVs (right panels). SVs are graphically represented as points (each point corresponds to an individual) coloured according to their positive/negative values.

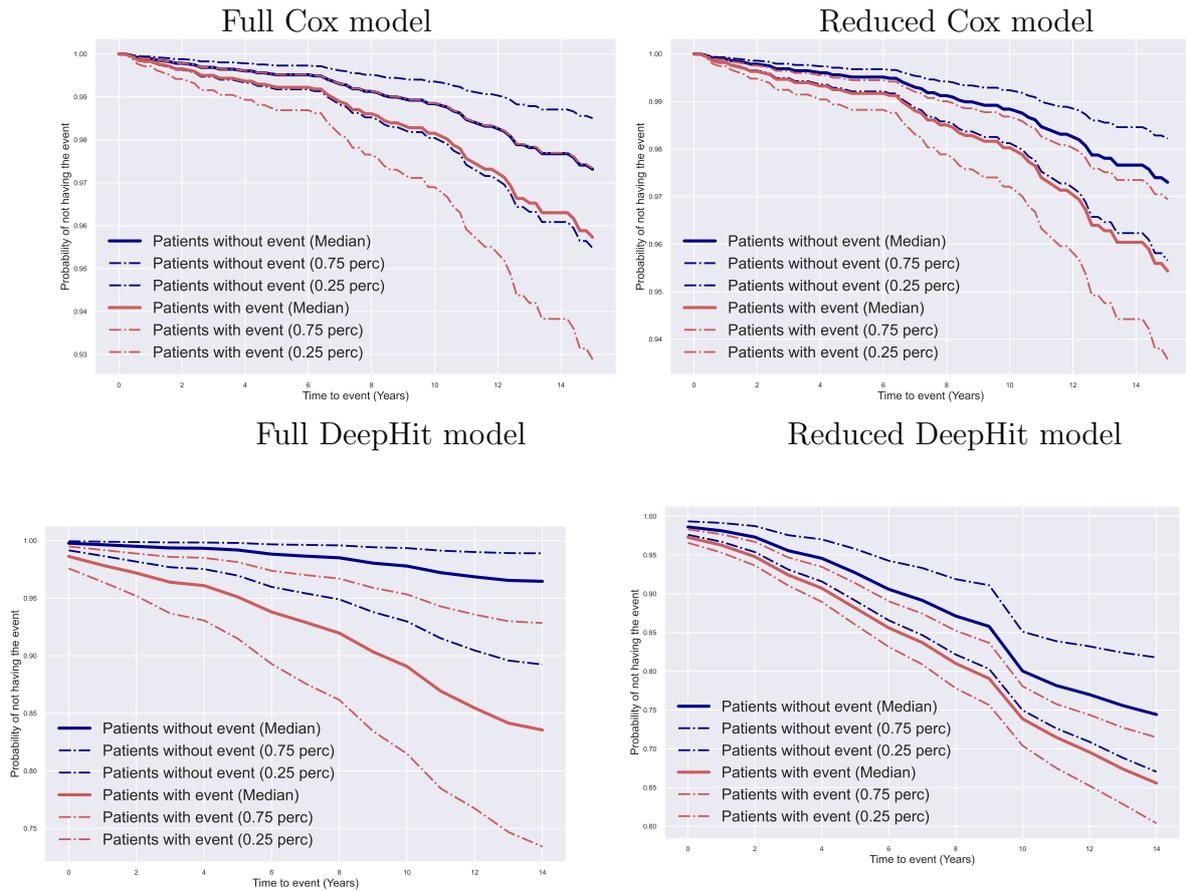


Figure 3. Survival curves distributions, stratified by patients with and without the event, estimated by the full Cox model (top left panel), the reduced Cox model (top right panel), the full DeepHit (bottom left panel) and the reduced DeepHit (bottom right panel), at the baseline.

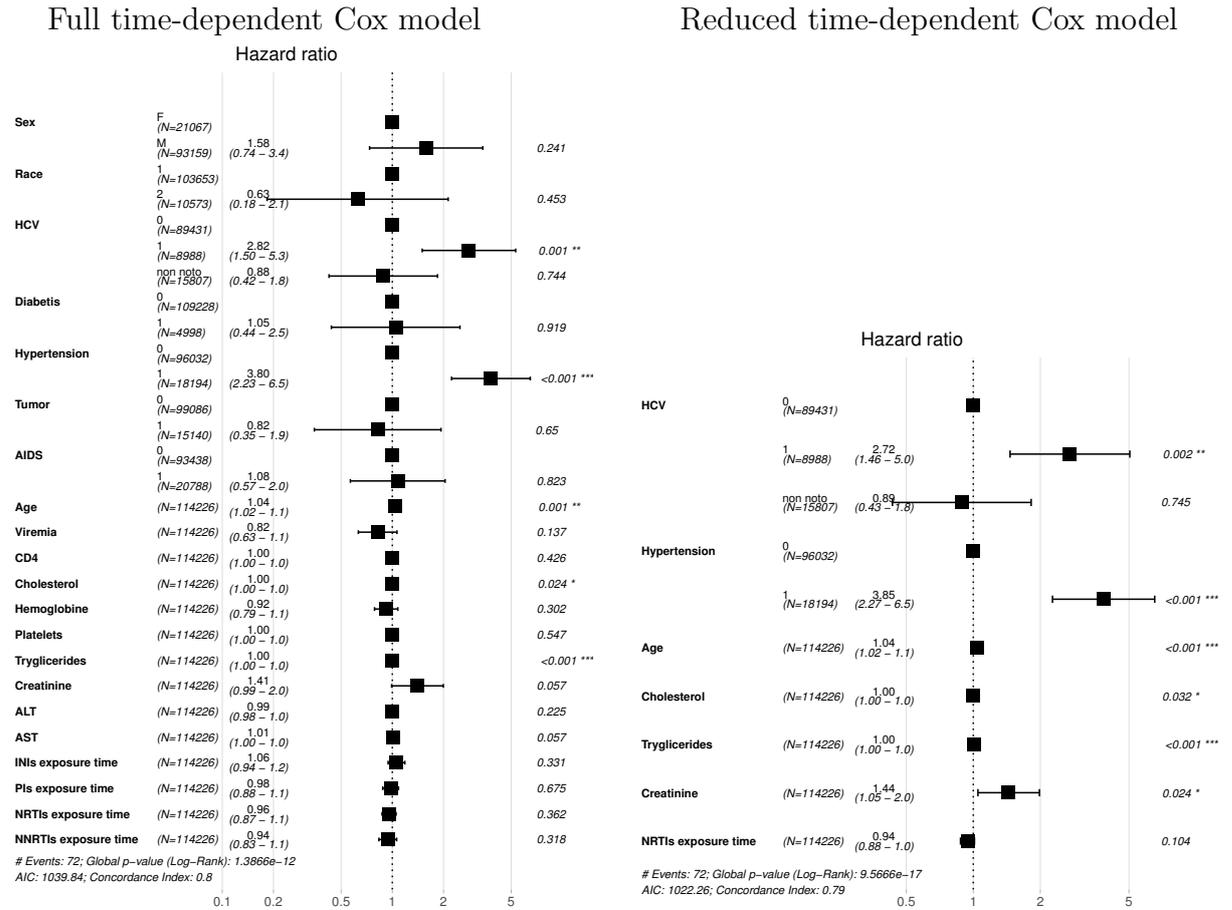


Figure 4. Results of the Cox time-dependent model fitted with the full (left panel) and the reduced (right panel) set of covariates, tracked over time.

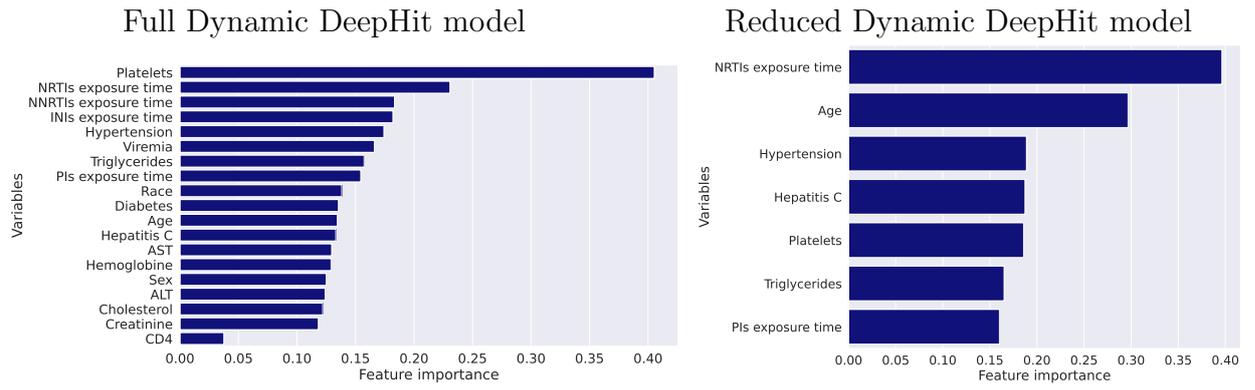


Figure 5. Features importance of the Dynamic DeepHit full (left panel) and reduced (right panel) model obtained with PFI.

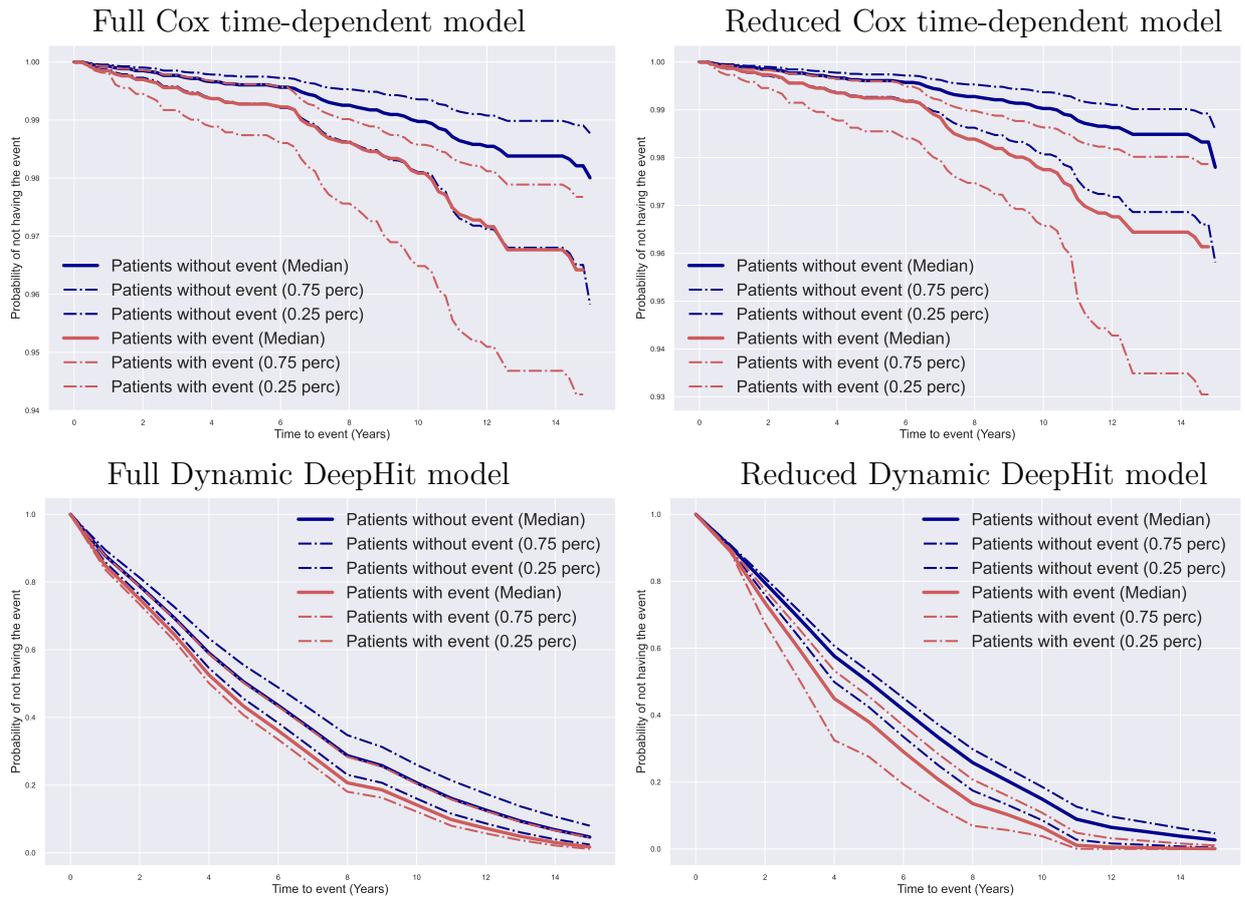


Figure 6. Survival curves distributions, stratified by patients with and without the event, estimated by the full and reduced Cox time-dependent models (top panels) and by the full and reduced Dynamic DeepHit models (bottom panels).

Name	Description	Type	Time-dependency
Sex	Gender of the patient	Binary (F, M)	Fixed
Race	Race of the patient	Binary (White, Other)	Fixed
HCV	Whether the patient has been diagnosed with hepatitis C at each visit	Categorical (Yes, No, Unknown)	Fixed
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
Age [Years]	Age of the patient updated at each visit	Continuous	Time-dependent
CD4 [$\frac{cells}{\mu L}$]	Level of CD4 at each visit	Continuous	Time-dependent
Cholesterol [$\frac{mg}{dL}$]	Level of cholesterol at each visit	Continuous	Time-dependent
Viremia [$\log_{10} \frac{copies}{mL}$]	Level of viremia at each visit	Continuous	Time-dependent
Creatinine [$\frac{mg}{dL}$]	Level of creatinine at each visit	Continuous	Time-dependent
Triglycerides [$\frac{mg}{dL}$]	Level of triglycerides at each visit	Continuous	Time-dependent
AST [$\frac{U}{L}$]	ASpartate aminoTransferase	Continuous	Time-dependent
ALT [$\frac{U}{L}$]	ALanine Transaminase	Continuous	Time-dependent
Platelets [$\frac{10^9}{L}$]	Number of platelets at each visit	Continuous	Time-dependent
Hemoglobin [$\frac{g}{dL}$]	Level of hemoglobin at each visit	Continuous	Time-dependent
PIs time [Years]	Cumulative years of protease inhibitor drug exposure	Continuous	Time-dependent
INIs time [Years]	Cumulative years of integrase inhibitor drug exposure	Continuous	Time-dependent
NRTIs time [Years]	Cumulative years of nucleotide reverse transcriptase inhibitor drug exposure	Continuous	Time-dependent
NNRTIs time [Years]	Cumulative years of non-nucleotide reverse transcriptase inhibitor drug exposure	Continuous	Time-dependent

Table 1

List and description of the 21 selected patient-specific covariates.

Statistics	Age		Viremia		CD4		Cholesterol		Hemoglobin		Platelets		Triglycerides	
	[Years]		[$\log_{10} \frac{copies}{mL}$]		[$\frac{cells}{\mu L}$]		[$\frac{mg}{dL}$]		[$\frac{g}{dL}$]		[$\frac{10^9}{L}$]		[$\frac{mg}{dL}$]	
-	Bl	End	Bl	End	Bl	End	Bl	End	Bl	End	Bl	End	Bl	End
Mean	38.484	48.283	3.789	0.941	423.519	739.5	170.093	184.489	14.194	14.924	217.93	234.1	130.266	129.455
Std	9.984	11.289	1.686	0.945	282.32	351.927	44.838	40.372	4.658	1.725	102.972	86.287	81.521	77.911
Min	0.38	17.99	0.278	0.0	0.0	1.0	0.5	37.0	6.5	6.5	3.0	2.0	0.88	18.0
25%	31.37	40.274	2.597	0.278	232.0	508.0	140.0	158.0	13.1	14.2	168.0	192.0	76.0	80.0
50%	37.496	48.327	4.338	0.278	373.5	706.5	166.0	182.0	14.4	15.2	210.0	229.0	107.0	107.0
75%	44.37	55.585	5.042	1.414	570.0	931.0	194.0	209.0	15.3	16.1	254.0	266.0	158.0	155.0
Max	82.962	91.206	7.053	6.471	2621.0	4595.0	927.0	452.0	157.0	19.8	800.0	800.0	500.0	500.0

Statistics	Creatinine		ALT		AST		Exp. time to INIs		Exp. time to PIs		Exp. time to NRTIs		Exp. time to NNRTIs	
	[$\frac{mg}{dL}$]		[$\frac{U}{L}$]		[$\frac{U}{L}$]		[Years]		[Years]		[Years]		[Years]	
-	Bl	End	Bl	End	Bl	End	Bl	End	Bl	End	Bl	End	Bl	End
Mean	0.848	1.05	43.951	33.932	36.595	31.957	-	2.047	-	2.844	-	7.17	-	2.85
Std	0.251	0.388	74.679	34.312	58.895	41.899	-	2.899	-	4.027	-	5.175	-	4.334
25%	0.0	0.22	2.0	4.0	5.0	4.0	-	0.0	-	0.0	-	0.0	-	0.0
25%	0.73	0.89	20.0	19.0	20.0	21.0	-	0.0	-	0.0	-	3.103	-	0.0
50%	0.83	1.02	29.0	25.0	26.0	25.0	-	0.736	-	0.171	-	6.544	-	0.0
75%	0.95	1.17	45.0	36.25	36.0	31.0	-	3.765	-	5.125	-	10.708	-	5.459
Max	8.73	9.89	2010.0	659.0	2592.0	1312.0	-	101.494	-	22.485	-	101.494	-	22.067

Statistics	Sex		Race		HCV		AIDS		Diabetes		Hypertension		Tumor	
	[F/M]		[White/Other]		[Y/N/Unknown]		[Y/N]		[Y/N]		[Y/N]		[Y/N]	
-	Bl	End	Bl	End	Bl	End	Bl	End	Bl	End	Bl	End	Bl	End
%	F = 20	F = 20	W = 91	W = 91	Y = 5	Y = 5	Y = 8	Y = 13	Y = 1	Y = 4	Y = 4	Y = 23	Y = 4	Y = 9
%	M = 80	M = 80	O = 9	O = 9	N = 73	N = 73	N = 92	N = 87	N = 99	N = 96	N = 96	N = 76	N = 96	N = 91
%	-	-	-	-	U = 22	U = 22	-	-	-	-	-	-	-	-

Table 2

Upper part of the table reports descriptive statistics of time-dependent numerical variables. Location indexes are reported for each variable both at the baseline (Bl) and at the end of the follow-up (End). Bottom part of the table reports descriptive statistics of the categorical variables: first three are time-invariant and last four are time-dependent and modelled as step functions. Percentages are reported both at baseline and at the end of the follow-up. Square brackets contain unit of measurements for numerical variables and classes for categorical ones.

Model	C-index training	C-index test	MSE (event)	Accuracy	Sensitivity	AUROC
Full Cox model	0.7816	0.6432	28.4786	0.6202	0.6667	0.66
Reduced Cox model	0.7643	0.6919	24.1500	0.6224	0.7222	0.71
Full DeepHit model	0.7513	0.6443	36.0501	0.7320	0.4444	0.72
Reduced DeepHit model	0.7642	0.6542	54.6865	0.6246	0.6667	0.72
Bootstrapped Cox model	0.7497	0.7853	29.2336	0.6892	0.7059	0.76
Bootstrapped DeepHit model	0.7627	0.7882	43.5907	0.6831	0.6765	0.75
Full TD Cox	0.8074	0.6854	26.3180	0.4983	0.7778	0.68
Reduced TD Cox	0.7914	0.7059	17.6951	0.3699	0.9444	0.71
Full Dynamic DeepHit	0.6954	0.7104	27.8194	0.7110	0.7778	0.83
Reduced Dynamic DeepHit	0.7025	0.6746	26.5441	0.7796	0.5556	0.78

Table 3
Models performances

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

- 81/2022** Bonizzoni, F.; Hauck, M.; Peterseim, D.
A reduced basis super-localized orthogonal decomposition for reaction-convection-diffusion problems
- 83/2022** Ciaramella, G.; Gander, M.; Mazzieri, I.
Unmapped tent pitching schemes by waveform relaxation
- 84/2022** Ciaramella, G.; Gambarini, M.; Miglio, E.
A preconditioner for free-surface hydrodynamics BEM
- 82/2022** Ciaramella, G.; Gander, M.; Van Criekingen, S.; Vanzan, T.
A PETSc Parallel Implementation of Substructured One- and Two-level Schwarz Methods
- 80/2022** Balduzzi, G.; Bonizzoni, F.; Tamellini, L.
Uncertainty quantification in timber-like beams using sparse grids: theory and examples with off-the-shelf software utilization
- 78/2022** Bucelli, M.; Gabriel, M. G.; Gigante, G.; Quarteroni, A.; Vergara, C.
A stable loosely-coupled scheme for cardiac electro-fluid-structure interaction
- 79/2022** Antonietti, P. F.; Farenga, N.; Manuzzi, E.; Martinelli, G.; Saverio, L.
Agglomeration of Polygonal Grids using Graph Neural Networks with applications to Multigrid solvers
- 77/2022** Ziarelli, G.; Dede', L.; Parolini, N.; Verani, M.; Quarteroni, A.
Optimized numerical solutions of SIRDVW multiage model controlling SARS-CoV-2 vaccine roll out: an application to the Italian scenario.
- 76/2022** Spreafico, M.; Ieva, F.; Fiocco, M.
Longitudinal Latent Overall Toxicity (LOTox) profiles in osteosarcoma: a new taxonomy based on latent Markov models
- 70/2022** Andrini, D.; Balbi, V.; Bevilacqua, G.; Lucci, G.; Pozzi, G.; Riccobelli, D.
Mathematical modelling of axonal cortex contractility