Journal of Mathematics and Statistics

Original Research Paper

Survival Analysis of Data of HIV Infected Persons Receiving Antiretroviral Therapy Using a Model-Based Binary Tree Approach

[1,2]Simon Tiendrébéogo, [1]Blaise Somé, [2]Séni Kouanda and [3,4]Simplice Dossou-Gbété

[1]Laboratoire d'Analyse Numérique, d'Informatique et de BIOmathématiques (LANIBIO), Université Joseph Ki-ZERBO, Ouagadougou, Burkina Faso
[2]Département Biomédical et Santé Publique, Institut de Recherche en Sciences de la Santé (IRSS), Ouagadougou, Burkina Faso
[3]Université de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France
[4]International Chair on Mathematical Physics and Applications ICMPA-Unesco Chair, UAC, Bénin

Article history
Received: 26-09-2019
Revised: 03-12-2019
Accepted: 21-12-2019

Corresponding Author:
Simon Tiendrébéogo
Laboratoire d'Analyse Numérique, d'Informatique et de BIOmathématiques, Université Joseph Ki-ZERBO, Ouagadougou, Burkina Faso.
Email: simon.tiendrebeogo@yahoo.fr

Abstract: Discrete-time approach is used in survival data analysis when only the time interval in which the event of interest has occurred is known or when this event occurs in a discrete - time scale. The work presented in this paper is motivated by the analysis of HIV/AIDS follow-up data collected in Burkina Faso during the 5-YEAR Global Fund program implemented to fight AIDS, Tuberculosis and Malaria. The research question that motivated the work is the likely existence of different mortality risk profiles of people infected with HIV/AIDS, depending on their characteristics and health status at the beginning of their care. In order to answer these questions, we considered a binary tree regression approach for survival data analysis since such a model owns the ability to handle interaction effects between the outcome covariates without a tight specification of such effects during the model statement step. This helps to prevent specification and interpretation errors. The fitted model resulted in splitting patients into three disjoint subgroups, corresponding each to a specific hazard profile.

Keywords: Model-Based Binary Regression Tree, Discrete Time-to-Event, Hazard Probability, Survival Analysis, HIV/AIDS, Antiretroviral Therapy

Introduction

From 2003 to 2007, the Global Fund supported health institutions in Burkina Faso to promote the access of HIV infected persons to Antiretroviral Therapy (ART), considered as a therapeutic advance in the fight against HIV/AIDS (Kouanda *et al*., 2008). Every six months, health data were recorded during clinical visits. An evaluation of this program was done in 2008 in order to assess the efficiency of the program and it involved the analysis of follow-up data gathered during the program execution. We sought to address two research questions in this paper: are there groups of patients with different and specific risk profiles? Which characteristics, among those that are recorded, are correlated with this risk and can help to predict accurately the hazard of death? To achieve this goal, we will use survival tree methods for the analysis of survival data. The main characteristic of this approach is its ability to capture interaction effects between predictors, specifically when there is a large number of predictors considered for modeling the distribution of a response variable. A binary tree is fitted to the dataset by recursively splitting covariates to create partitions of covariate space in order to obtain homogeneous groups with respect to the studied response.

Contributions in tree-based methods for discrete-time survival analysis include (Bou-Hamad *et al*., 2009) and (Schmid *et al*., 2016). Both methods consider that time-to-event data are observed jointly with covariates that describe individuals and consider that hazard of event occurrence is a function of time and covariates. Let us consider $X = (X_1, X_2,\ldots, X_p)$ a vector of covariates, and $\{N_k, k = 1,\ldots, K\}$ a partition of the covariates space denoted by $\mathcal{D}(X)$. Let h denote the hazard function. In Bou-Hamad *et al*. (2009) approach, hazard probability has been modeled as follows: for all $x \in \mathcal{D}(X)$ and time index t,

$$\log\left(\frac{h(t \mid x)}{1-h(t \mid x)}\right) = \alpha(t \mid x) \text{ with } \alpha(t \mid x) = \sum_{k=1}^{K} \alpha_k(t) I\left(x \in N_k\right) \quad (1)$$

where, $I(x \in A)$ is the indicator function of the set A. In the Schmid *et al*. (2016) approach, the time variable denoted T is a candidate splitting variable for tree construction. Schmid *et al*. (2016) consider a partition $\{N_k, k = 1,\ldots, K\}$ of the input space $\mathcal{D}(T,X)$ and a hazard model specified by:

$$\log\left(\frac{h(t \mid x)}{1-h(t \mid x)}\right) = \sum_{k=1}^{K} \alpha_k (t,x) I\left((t,x) \in N_k\right) \quad (2)$$

where t is a time index and x is an observed value of the covariate vector X.

Our present work proposes a semiparametric model-based binary tree for the analysis of the follow-up data of the HIV infected people who were included in the Global Fund program, with the aim of analyzing the correlation between the survival of the people who had access to this therapy, and their health status at the time of admission to the program, as well as the means by which these patients arrived at the program, the characteristics of the patients and the attributes of care facilities. Unlike (Bou-Hamad *et al*., 2009) and (Schmid *et al*., 2016) methods, the proposed approach distinguishes two groups of covariates: the first group is involved in the binary tree construction, and the second group is involved in the statement of a parametric regression model in each node of the tree.

The rest of the paper unfolds as follows. Section 2 is devoted to the exposition of the statistical model proposed for discrete-time survival analysis and the model fitting algorithm; in section 3, survival data are analysed on the basis of the fitted model; section 4 is devoted to the discussion of the results.

Statistical Methods

Notations

Let us consider a sequence of time intervals $[0,t_1[,[t_1,t_2[,\ldots,[t_{s-1},t_s[,\ldots$ where the subscript s indexes a time interval. The covariates are splited into two parts. The first one is made of covariates $Z_1, Z_2,\ldots, Z_q$, called moderators according to (Fokkema *et al*., 2018) and will be involved in the binary tree construction. These covariates are used to characterize subgroups of the population through a partition of the input space $\mathcal{D}(Z) = \prod_{j=1}^{q} \mathcal{D}(Z_j)$ where $Z = (Z_1, Z_2,\ldots, Z_q)$ and $\mathcal{D}(Z_j)$ denotes the set of all possible outcomes of Z_j, $j = 1,\ldots, q$. The second set of covariates $X_1, X_2,\ldots,X_p$, is involved in the statement of a parametric regression model, the linear prediction of a generalized linear model (glm). Let $X = (X_1, X_2,\ldots, X_p)$. We consider $T_E = 1, 2,\ldots$ and $T_C = 1, 2,\ldots$, a random event time interval index, and a random censoring time interval index respectively. Let $T := \min(T_E, T_C)$. We denote by $\Delta := I(T_E \leq T_C)$ the random binary variable that indicates whether T_E, the time to event, is censored[1] or not. The observed sample is denoted by $\{(t_i, \delta_i, x_i, z_i), i = 1,\ldots, n\}$ where t_i is the observed time of individual i, δ_i indicating whether the individual is censored or not, $z_i = (z_{1i}, z_{2i},\ldots, z_{qi})$, $x_i = (x_{1i},x_{2i},\ldots, x_{pi})$ are the observed values of Z and X.

Binary Regression Tree Model for Discrete-Time Hazard

The discrete-time hazard function h is the conditional probability that a randomly selected individual will experience death in time interval s, given that he didn't experience death prior to s. For a time index s, the discrete hazard is defined by:

$$h(s \mid x,z) = \Pr\left(T_E = s \mid T_E \geq s, X = x, Z = z\right) \quad (3)$$

where X is the vector of the predictors, Z is the vector of the moderators, T_E is the time-to-event variable and s is the observed time. We consider a discrete hazard model specified by:

$$g\left(h(s \mid x,z)\right) = \alpha(s \mid z) + \sum_{l=1}^{p} \beta_l (z) x_l \quad (4)$$

where s is a time index, x is an observed value of the covariate vector X, z is an observed value of the moderator vector Z and g is a strictly monotically increasing link function. It clearly appears that the model is a varying-coefficient model (Hastie and Tibshirani, 1993): coefficients α_s and β_l change with the value of moderator variables Z_m, $m = 1,\ldots, q$. Those variables, called effect modifiers by (Hastie and Tibshirani, 1993) are distinct from covariates X_l, $l = 1,\ldots, p$ involved in the linear model. Given a partition $\{N_k, k = 1,\ldots, K\}$ of the moderators space $\mathcal{D}(Z)$, let us consider:

$$\alpha(s \mid z) = \sum_{k=1}^{K} \alpha_k (s) I\left(z \in N_k\right) \text{ and } \beta_1 (z) = \sum_{k=1}^{K} \beta_{lK} I\left(z \in N_k\right).$$

Therefore:

$$g\left(h(s \mid x,z)\right) = \sum_{k=1}^{K} \alpha_k (s) I\left(z \in N_k\right) + \sum_{k=1}^{K}\sum_{l=1}^{p} \beta_{lK} I\left(z \in N_k\right) x_l . \quad (5)$$

It turns out that:

$$h(s \mid x,z) = \sum_{k=1}^{K} g^{-1}\left(\alpha_k (s) + \sum_{l=1}^{p} \beta_{lK} x_1\right) I\left(z \in N_k\right). \quad (6)$$

Let $h_k(s \mid x) := g^{-1}\left(\alpha_k (s) + \sum_{l=1}^{p} \beta_{lK} x_1\right)$.

Then $h(s \mid x,z) = \sum_{k=1}^{K} h_k (s \mid x) I\left(z \in N_k\right)$.

[1] A patient is censored if he is lost sight of after his last follow-up visit

When there are no covariates but only moderators, Equation (4) reduces to:

$$g\left(h(s\mid z)\right)=\alpha(s\mid z)$$
$$=\sum_{k=1}^{K}\alpha_k(s)I\left(z\in N_k\right) \quad (7)$$

which is the model considered by Bou-Hamad *et al*. (2009). It is claimed in (Singer and Willett, 1993) that the specification of time effect as $\alpha_k(s)$ in (6) is the most general parameterization of time effect. Classical link functions include probit, logit and cloglog functions. The latter leads to discrete-time counterpart of an extended Cox proportional hazard model with respect to a link function $f(h(s|x)) = -\log(1-h(s \mid x)) = \exp(\alpha(s))\exp(\beta x)$ where $\exp(\alpha(s))$ stands for the baseline hazard.

Augmented Design Matrix

For any individual i, let t_i be the index of the time interval where i experienced the event or was censored and $y_{it_i} = \delta_i$. Let us define sequences $y_i = (0,0,\ldots, y_{it_i}) = (0,0,\ldots, 1)$ if i is uncensored and $y_i = (0,0,\ldots, y_{it_i}) = (0,0,\ldots, 0)$ if i is censored. The length of sequence y_i is t_i. The tree construction is based on moderators. The likelihood of model (4) is given by:

$$L=\prod_{i=1}^{n}\prod_{s=1}^{t_i}\left[h\left(s\mid x_i,z_i\right)\right]^{y_{is}}\left[1-h\left(s\mid x_i,z_i\right)\right]^{1-y_{is}}. \quad (8)$$

One can consider (8) as the likelihood of a binomial model with probabilities $h(s \mid x_i, z_i)$ and independent observations y_{is} of independent statistical binary variables y_{is}. It turns out that the model parameters can be estimated by using binary response regression techniques. For that purpose, we need to create an (augmented) data matrix (Berger and Schmid, 2017) with t_i rows derived from the initial data $\{(t_i, \delta_i, x_i, z_i), i = 1,\ldots, n\}$. In this matrix, the sth row contains information about the sth time interval. The first column of the data matrix is related to Δ, the second is related to T (observed time), the p subsequent columns consist of the observed values of X and the last q columns are the observed values of Z. The rows of the augmented design matrix corresponding to a subject i that has experienced the event ($\delta_i = 1$) are given by:

$$M_{i1}=\begin{array}{c} \begin{matrix} Y & S & X_1 & \cdots & X_p & Z_1 & \cdots & Z_q \end{matrix} \\ \begin{pmatrix} 0 & 1 & x_{1i} & \cdots & x_{pi} & z_{1i} & \cdots & z_{qi} \\ 0 & 2 & x_{1i} & \cdots & x_{pi} & z_{1i} & \cdots & z_{qi} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & t_i-1 & x_{1i} & \cdots & x_{pi} & z_{1i} & \cdots & z_{qi} \\ 1 & t_i & x_{1i} & \cdots & x_{pi} & z_{1i} & \cdots & z_{qi} \end{pmatrix} \end{array}$$

In the case of a censored subject i ($\delta_i = 0$), the rows corresponding to i are given by:

$$M_{i0}=\begin{array}{c} \begin{matrix} Y & S & X_1 & \cdots & X_p & Z_1 & \cdots & Z_q \end{matrix} \\ \begin{pmatrix} 0 & 1 & x_{1i} & \cdots & x_{pi} & z_{1i} & \cdots & z_{qi} \\ 0 & 2 & x_{1i} & \cdots & x_{pi} & z_{1i} & \cdots & z_{qi} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & t_i-1 & x_{1i} & \cdots & x_{pi} & z_{1i} & \cdots & z_{qi} \\ 0 & t_i & x_{1i} & \cdots & x_{pi} & z_{1i} & \cdots & z_{qi} \end{pmatrix} \end{array}$$

By stacking these augmented matrices one obtains a design matrix that will be used by the model fitting algorithm. Singer and Willet (1993) proposed a similar design matrix. In short, the difference is on the Time variable. Instead of one variable, the Time variable is represented by $t_{\max}$ time indicators $\left\{D_1, D_2, \ldots, D_{t_{\max}}\right\}$ with $t_{\max}$ the maximum length of time an individual was alive.

Model Fitting

Model-Based recursive partitioning (MOB) is an algorithm that aims to produce unbiased binary tree where fitted parametric models are associated with each terminal node (Zeileis *et al*., 2008; Hothorn *et al*., 2006). We have used MOB to build the tree. The algorithm finds the partitions after completion of four steps (Algorithm 1). More details on parameter instability tests can be found in (Hornik and Zeileis, 2007; Hothorn *et al*., 2006). Unlike other binary regression tree methods, MOB does not require post-pruning to avoid overfitting and it results in an optimally sized tree. For tree construction, two hyperparameters are provided to run the algorithm: the significance level for parameter instability tests denoted α and the minimum terminal node size denoted minsize. The terminal node size refers to rows of the augmented data matrix. The final tree is obtained by the selection of an optimal value among hyperparameter values. This can be achieved by looking for an optimal joint value of a minimal terminal node size minsize and a numeric significance level α by using Bayesian Information Criterion (BIC).

Algorithm 1 MOB Algorithm

Step 1: **Fit the model to the dataset**.
Let us consider the augmented data matrix previously described to be the dataset. It consists of two sets of variables: model variables and moderator variables. The model can be fitted by maximizing the log-likelihood resulting from (8). We denote by $\hat{\theta}$ the parameter estimate.

Step 2: **Test for parameter instability with respect to every partitioning variable**.
Let τ be the current node and $\hat{\theta}_\tau$ the estimate of parameter θ_τ in τ. We denote by τ_L and τ_R the children nodes resulting from a binary partition of τ. The estimate $\hat{\theta}_\tau$ is said to be unstable if there is a moderator covariate Z_j, $j = 1: q$ such that $\hat{\theta}_{\tau_L}\left(z_j\right)$ and

$\hat{\theta}_{\tau_R}(z_j)$ respectively the parameter estimates in τ_L and τ_R are significantly different. Generalized M-fluctuation tests are used for that purpose (Hornik and Zeileis, 2007). The sup LM statistic is used for numerical moderators and a χ^2 statistic is used for categorical moderators (Hornik and Zeileis, 2007). See supplementary material for details on sup LM statistic.

Step 3: **If there is some overall parameter instability, split the dataset with respect to the variable associated with the highest instability (the smallest *p*-value) into two children nodes**.
To determine whether there is some overall instability, it is checked whether the minimal *p*-value falls below α, a pre-specified significance level. The Bonferroni method can be used to adjust for multiple testing. The split point is found by applying an exhaustive search procedure: for every conceivable split point, the parametric model is fitted in each one of the two children nodes generated by this split point and then the split associated with the maximum sum of the two observed log-likelihoods in the children nodes is chosen.

Step 4: **Repeat the procedure in each of the resulting subgroups until no significant instability is detected or a minimum terminal node size criterion is met**.

This method allowed us to cover a large set of hyperparameter values and then to select the optimal tree using the BIC criterion. We used the partykit package (Hothorn and Zeileis, 2016) to fit the model. Hazard probabilities were estimated using (6).

Stability Analysis

The success of binary trees among statistical methods of decision-making should not obscure the potential instability of models resulting from the execution of the algorithm used to fit the data. Therefore, it is essential to ensure the stability of the fitted model before its subsequent use, as for prediction task. Stability assessment can be done by fitting the same model to bootstrapped samples from the training dataset. Bootstrap trees may select variables and cutpoints that were not selected by the original tree. Metrics for stability assessment include the relative variable selection frequency, the mean frequency of the variable selections per tree and the frequency of each cutpoint over the trees (Philipp *et al*., 2016):

- The relative variable selection frequency for a splitting variable z_j, $j = 1,\ldots, q$ equals the total number of bootstrap trees that have selected z_j at least once, divided by the total number of bootstrap trees.
- The mean frequency of the variable selections per tree for z_j is the total number of times z_j is selected for splitting by a bootstrap tree over the repetitions, divided by the total number of bootstrap trees.
- The relative frequency of a cutpoint $c(z_j)$ equals the total number of bootstrap trees that have selected $c(z_j)$ to split the variable z_j, divided by the total number of bootstrap trees.

A variable selection is stable if its frequency of selection is close to 100% and its average split count is close to its number of selections in the original tree. Different graphics are used to hightlight a variable cutpoint variability depending on the nature of the variable (categorical, numerical). For an ordered categorical variable, a barplot is used to show the frequency of all possible cutpoints. A histogram is used to illustrate the cutpoint variability when the splitting variable is numerical. It is expected that the cutpoints selected in the original tree have the highest frequencies (one or more peaks in the histogram). For an unordered categorical variable, a specific plot is used to visualize the partitions' variability over the repetitions. The plot uses the same color for categories that belong to the same node. The combination of categories that corresponds to a partition observed in the original tree is marked on the right side of the plot by a solid red line. In addition, two dashed lines enclose the area representing the partition. The level(*s*) of the corresponding split(*s*) in the original tree are indicated by the number(s) on the right side of the area. To sum up, a cutpoint is stable if it is selected by most resampling trees. More details on the approach can be found in (Philipp *et al*., 2016).

We resort to a semiparametric bootstrap method which consists of two steps: a sampling step and an assignment step. During the first step, we sample with replacement from the survival data (t_i, δ_i), $i = 1,\ldots, n$. In the second step, we assign covariate vectors x_i, z_i conditional on the (t_i, δ_i) sampled on the basis of a probability distribution determined by the fitted model (Zelterman *et al*., 1996). The fitted hazard probability is used for uncensored individuals ($\delta_i = 1$) and the fitted survival probability is used for censored individuals ($\delta_i = 0$).

Data Analysis

Study Population and Measurements

The population concerned by the study is the overall HIV infected patients that had started antiretroviral therapy in Burkina Faso between 1st January 2003 and 31st December 2007. The country health system is splitted into 13 health regions totalizing 55 health districts and 77 ART centers. Among other initiatives, the ART centers had received funding from the Global Fund to fight AIDS, Tuberculosis and Malaria. For the purpose of the Global 5-Year program evaluation, those centers were classified as low, medium and high scale, according to the level of funding received by health districts in 2007. In each ART center, all the patients older than 15 that had laboratory documentation of HIV infection and that had received Highly Active Antiretroviral Therapy (HAART) in the

center for at least 6 months of follow-up were included in the data collection. Among the 5608 patients initially studied by (Kouanda *et al*., 2011), 1267 patients with missing initial CD4 count or unknown WHO clinical stage were excluded for the present analysis. People who had begun ART before they joined the ART centers involved in the evaluation were not included in this study. We didn't consider the covariates Body Mass Index (BMI) and time-dependent CD4 count because of the large proportion of missing values.

Data Description

4341 patients were included in our analyses: 70.6% were Female and 71.0% were under 40 years old (Table 1). The data are made of five cohorts: among the patients, 4.9% started HAART in 2003, 16% in 2004, 26.0% in 2005, 29.0% in 2006 and 24.1% in 2007. At initiation, most of the patients had advanced HIV infection: 83.3% had started ART with CD4 count < 200 cells/μL and 80.5% were at WHO clinical stages III or IV. In 95% of all cases, the treatment regimen was two Nucleoside Reverse Transcriptase Inhibitors (NRTI), plus one non-Nucleoside Reverse Transcriptase Inhibitor (NNRTI). At the date of collection, 77.5% were alive, 7.0% were lost-to-follow-up and 11.6% were dead. The empirical hazard function decreased from 7.49% in semester 1 to 2.02% in semester 2 (Table 2).

Table 1: Characteristics of patients at the time of starting treatment

Variable (acronym)	2003 (n = 215) n (%)	2004 (n = 681) n (%)	2005 (n = 1127) n (%)	2006 (n = 1274) n (%)	2007 (n = 1044) n (%)	Total (n = 4341) n (%)
Gender (Genre)						
Female	157 (73.0)	485 (71.2)	808 (71.7)	878(68.9)	739 (70.8)	3067 (70.6)
Male	58 (27.0)	196 (28.8)	319 (28.3)	396 (31.1)	305 (29.2)	1274 (29.4)
Age (Age)						
15-29	58 (27.0)	158 (23.2)	274 (24.3)	298 (23.4)	268 (25.7)	1056 (24.3)
30-39	109 (50.7)	345 (50.7)	544 (48.3)	575 (45.1)	452 (43.3)	2025 (46.7)
>= 40	48 (22.3)	178 (26.1)	309 (27.4)	401 (31.5)	324 (31.0)	1260 (29.0)
HIV type (Serologie)						
HIV1	206 (95.8)	650 (95.5)	1063 (94.3)	1210 (95.0)	993 (95.1)	4122 (95.0)
Others	9 (4.2)	31 (4.5)	64 (5.7)	64 (5.0)	51(4.9)	219 (5.0)
CD4 count (inCD4)						
< 50	38 (17.7)	126 (18.5)	289 (25.6)	256 (20.0)	189 (18.0)	898 (20.7)
50-99	61 (28.4)	146 (21.4)	213 (18.9)	257 (20.2)	181 (17.4)	858 (19.8)
100-199	85 (39.5)	265 (38.9)	470 (41.7)	564 (44.3)	478 (45.8)	1862 (42.9)
>=200	31 (14.4)	144 (21.2)	155 (13.8)	197 (15.5)	196 (18.8)	723 (16.6)
WHO clinical stage (StadeOMS)						
WHO stage I or II	68 (31.6)	112 (16.4)	199 (17.7)	250 (19.6)	216 (20.7)	845 (19.5)
WHO stage III	96 (44.6)	325 (47.7)	639 (56.7)	705 (55.3)	575 (55.1)	2340 (53.9)
WHO stage IV	51 (23.7)	244 (35.8)	289 (25.6)	319 (25.0)	253 (24.2)	1156 (26.6)
Outcome (death)						
Censored	187 (87.0)	585 (85.9)	960 (85.2)	1142 (89.6)	961 (92.0)	3835 (88.3)
Dead	28 (13.0)	96 (14.1)	167 (14.8)	132 (10.4)	83 (8.0)	506 (11.7)
Entry mode (EntryMod)						
NGO	30 (13.9)	85 (12.5)	184 (16.3)	191 (15.0)	174 (16.7)	664 (15.3)
Health facilities	111 (51.6)	467 (68.6)	718 (63.7)	818 (64.2)	701 (67.1)	2815 (64.8)
Relatives	25 (11.6)	25 (3.7)	68 (6.0)	136 (10.7)	67 (6.4)	321 (7.4)
Transfer	49 (22.8)	104 (15.3)	157 (13.9)	129 (10.1)	102 (9.8)	541 (12.5)
Health District (District)						
Bogodogo	92 (42.8)	176 (25.8)	261 (23.2)	229 (18.0)	229 (16.7)	932 (21.5)
Boulmiougou	118 (54.9)	492 (72.2)	693 (61.5)	747 (58.6)	613 (58.7)	2663 (61.3)
Others	5 (2.3)	13 (1.9)	173 (15.3)	298 (23.4)	257 (24.6)	746 (17.2)
Intensity of intervention (Scale)						
low scale	3 (1.4)	9 (1.3)	95 (8.4)	107 (8.4)	81 (7.8)	295 (6.8)
medium scale	31 (14.4)	18 (2.6)	146 (12.9)	286 (22.4)	231 (22.13)	712 (16.4)
high scale	181 (84.2)	654 (96.0)	886 (78.6)	881 (69.1)	732 (70.1)	3334 (76.8)

Note: NGO, Non-governmental organization

Table 2: Empirical risks table

Time index	Number at risk	Number of deaths	Number of dropouts	Hazard (%)
1	4341	325	202	7.49
2	3814	77	370	2.02
3	3367	42	669	1.25
4	2656	25	542	0.94
5	2089	17	517	0.81
6	1555	10	499	0.64
7	1046	10	1036	0.96

For statistical analysis, time indexes 7, 8, 9, 10 and 11 have been grouped into a single category 7. For the model building, we consider the variables Gender (Genre), Age, Entry mode (EntryMod) that describes the patients as predictors in the linear model. Potential partitioning variables were baseline CD4 count (inCD4), WHO clinical stage (StadeOMS), Intensity of the intervention (Scale) and health district category (District).

Results

Identified Subgroups and Hazards of Death Profiles

The fitted tree results in three terminal nodes (Fig. 1). The WHO clinical stage (StadeOMS) is selected for the first split, revealing that among the predictors, the baseline disease stage is the most closely correlated with the mortality hazard trajectories. Within the group of patients with one of the first three WHO clinical stages, the baseline CD4 count (inCD4) is the most correlated with the mortality hazard trajectories and induces a new split of the group. CD4 count is known to be a good predictor of the HIV dynamics during treatment in resource-limited countries. Patients with baseline CD4 count $\leq$ 50 cells/μL have a risk profile that is different from that of patients with baseline CD4 count > 50 cells/μL. Figures 2 to 4 illustrate the correlation between covariates and the hazard function. In the subgroup of patients with WHO clinical stage 4, the hazard function highly decreases from semester 1 to semester 2 (Fig. 2). The hazards of death in semester 1 are slightly higher for male patients, compared to female patients in all categories of patients and higher for 30-40 age group compared to the other age groups. But there is no significant difference in hazards of death between the categories of patients from the semester 2.

In the subgroup of patients with baseline CD4 count $\leq$ 50 cells/μL and WHO clinical stage $\leq$ 3, patients supported by NGOs or Relatives have similar hazard profiles and the lowest hazard estimates in semester 1 (Fig. 3). Transferred patients have the highest hazard estimates in semester 1. Patients aged 40 and over have the highest hazard estimates. The difference in risk between age categories is lower in patients followed by NGOs or supported by relatives than in transferred patients. In addition, the hazard function increases between semester 6 and time interval 7. The increase is lower in patients followed by NGOs or supported by relatives compared to patients from other modes of entry.

In the subgroup of patients with baseline CD4 count > 50 cells/μL and one of the first three WHO clinical stages (Fig. 4), the hazards of death in semester 1 are slightly higher for male patients compared to female patients and significantly higher for transferred patients compared to patients from other modes of entry. The hazard estimates are also higher for patients between the ages of 30 and 40. Patients from other age categories have similar hazard profiles. For all categories, the hazard function remains constant after the third semester.

A comparison of the results from the three subgroups reveals main features. First, in all subgroups, the hazard function decreases significantly between semester 1 and semester 2. The hazard estimates during these two semesters are higher in the subgroup of patients with baseline WHO clinical stage 4, compared to the two other subgroups of patients. The subgroup with baseline CD4 count > 50 cells/μL and one of the first three disease stages has the lowest hazard estimates during these two semesters. These findings underline, on the one hand the treatment efficacy for all HIV infected persons and on the other hand, that the efficacy is best for patients that initiate treatment at an early stage of infection. Secondly, hazard profiles differ significantly depending on how patients are entered into the active list of persons living with HIV/AIDS under ART. Patients supported by parents or NGOs have similar risk profiles. Patients recruited through health facilities and transfers also have comparable profiles. Thirdly, except in the subgroup of patients with baseline CD4 count $\leq$ 50 cells/μL and WHO clinical stage $\leq$ 3, hazard estimates in semester 1 are higher for male patients compared to female patients. Lastly, for all subgroups, there is a difference in hazard between age categories in the first semester. This difference is well illustrated in the subgroup of patients with baseline CD4 count > 50 cells/μL and one of the first three disease stages (Fig. 4).

Assessment of the Fitted Binary Tree Stability

We used 500 bootstrap samples for the stability analysis. Figure 5 and Table 3 show that the relative frequency of selecting WHO clinical stage (StadeOMS on the table), was 100%. In addition, Fig. 7 shows that

all the bootstrap trees have splitted StadeOMS at stage ≤ 3 on the first level as in the original tree. Thus, the variable WHO clinical stage can be considered as stable. For the variable inCD4, the relative frequency of selection is also 100% but, unlike in the original tree, the variable is selected 1.9 times for splitting by each bootstrap tree (Table 3). Most bootstrapped trees have selected the same cutpoint as in the original tree (CD4 count ≤ 50 cells/μL) on the second level. In most cases, an additional cutpoint (between 80 and 200) is also selected. This second cutpoint indicates that the subgroup of patients with CD4 count > 50 cells/μL may be splitted into two categories with specific survival profiles by a cutpoint between 80 cells/μL and 200 cells/μL. The upper bound of the possible cutpoints is known to be a threshold limit under which a HIV infected person is immunodepressed and in very urgent need of treatment. To sum up, the variable CD4 count is definitely relevant for predicting survival of ART patients although its split is less stable than the split of the variable WHO clinical stage. About 69.8% of the trees were built by selecting only the variables WHO clinical stage and baseline CD4 count as in the original tree (Fig. 6).

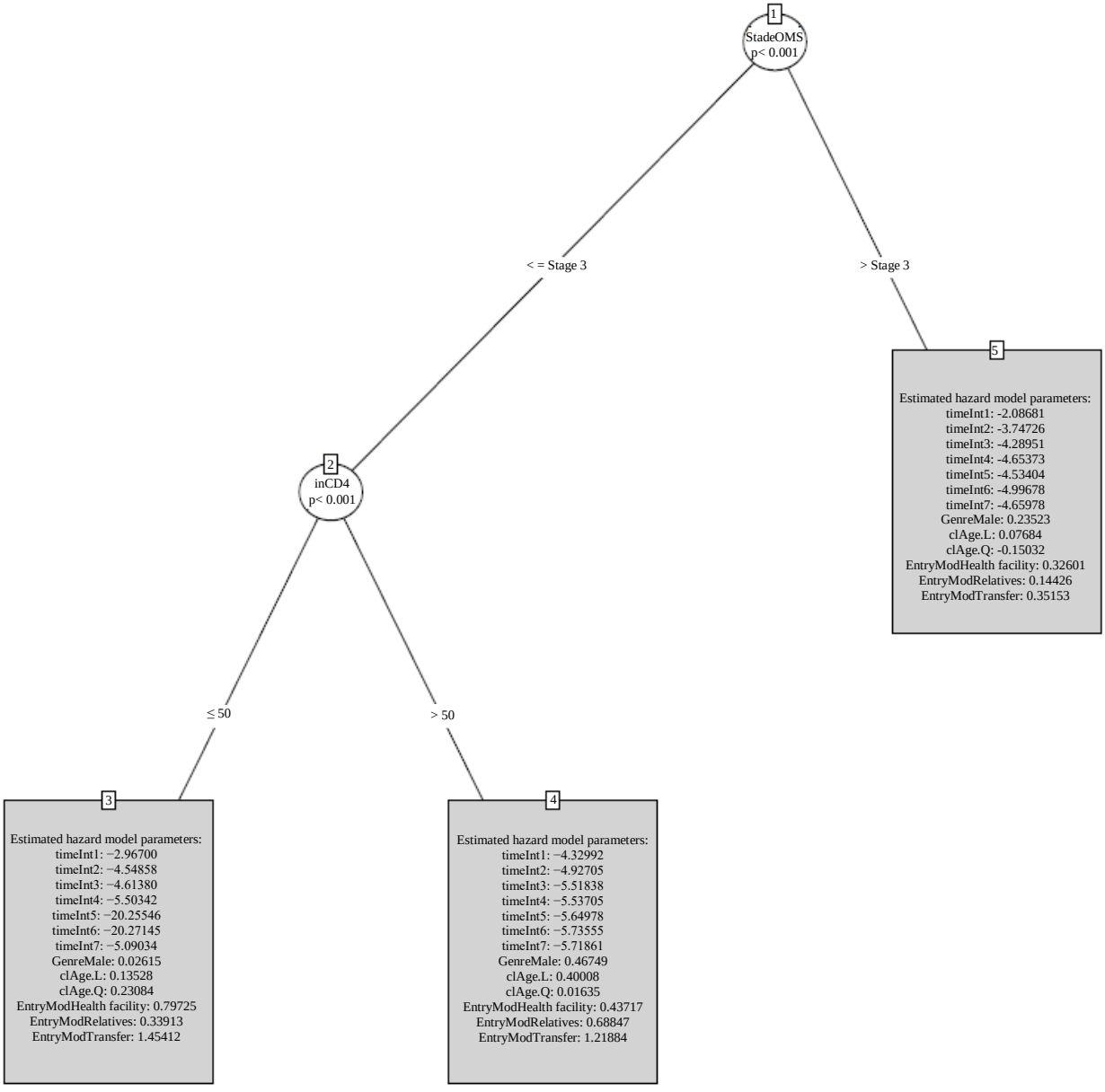

Fig. 1: Estimated discrete-time cloglog - hazard tree. The optimal hyperparameters (determined by BIC criterion) for model fitting are α = 0.01 and minsize = 1950

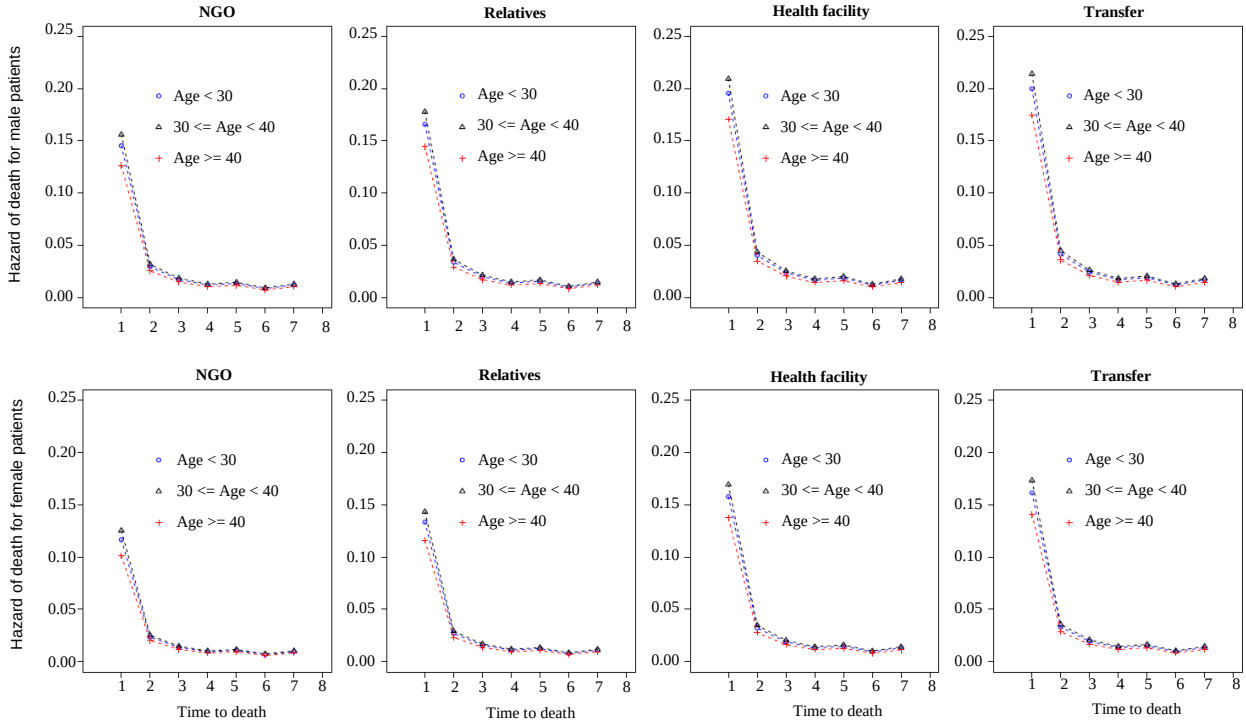

Fig. 2: Mortality hazard trajectories for different categories of patients that exist in the subgroup of patients with WHO clinical stage 4. At the top of the figure, hazard of deaths for male patients are depicted. At the bottom, those for female patients are illustrated

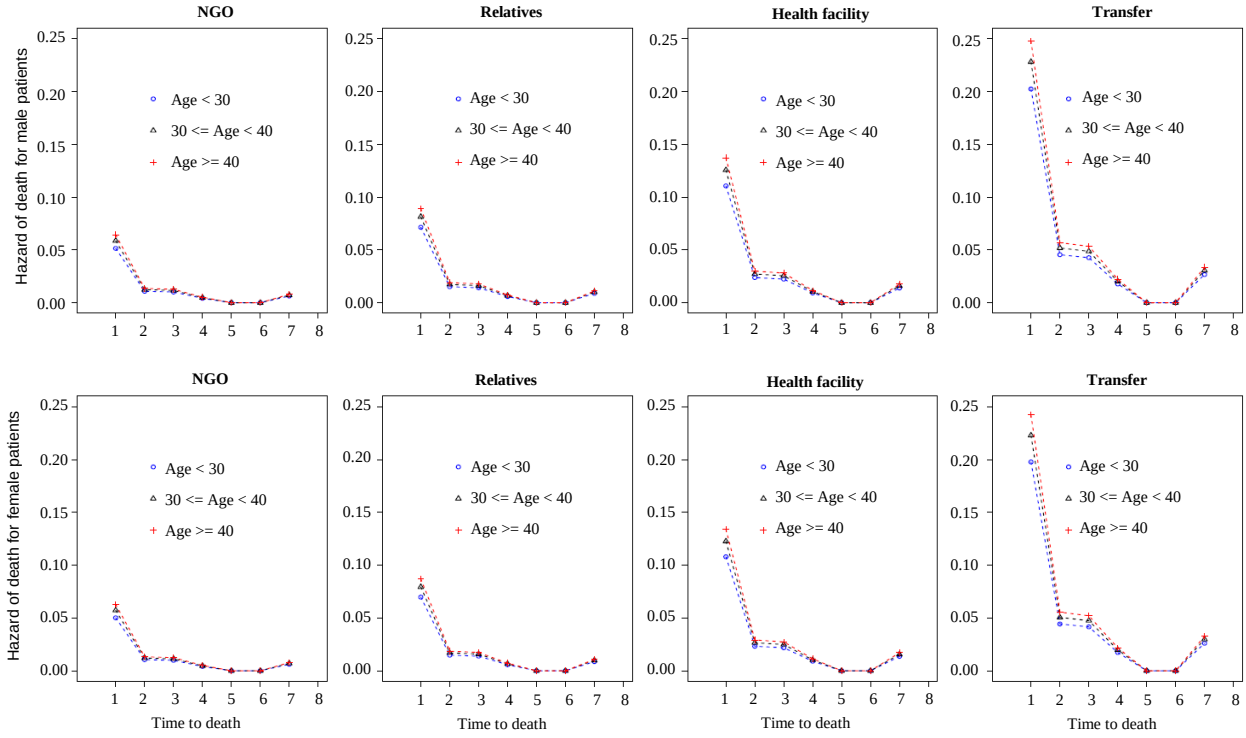

Fig. 3: Mortality hazard trajectories for different categories of patients that exist in the subgroup of patients with baseline CD4 count $\leq$ 50 cells/μL and WHO clinical stage $\leq$ 3. At the top of the figure, hazard of deaths for male patients are depicted. At the bottom, those for female patients are illustrated

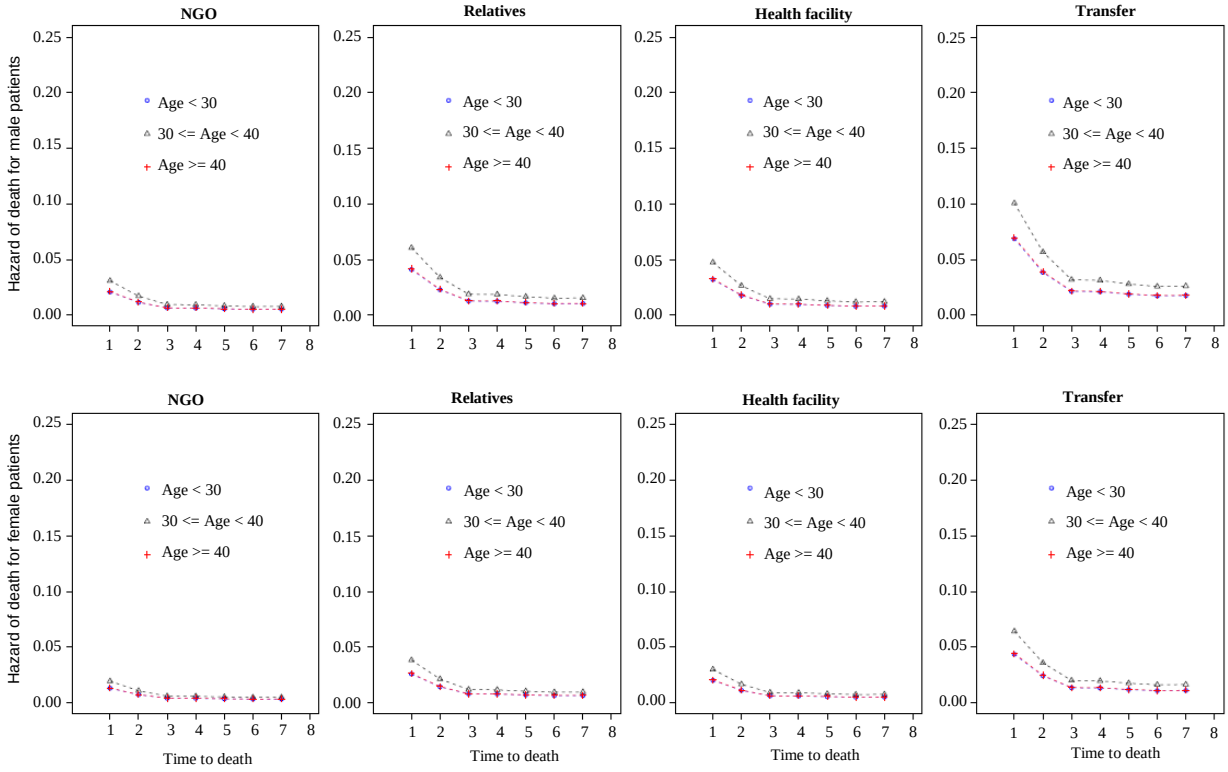

Fig. 4: Mortality hazard trajectories for different categories of patients that exist in the subgroup of patients with baseline CD4 count > 50 cells/μL and WHO clinical stage ≤ 3. At the top of the figure, hazard of deaths for male patients are depicted. At the bottom, those for female patients are illustrated

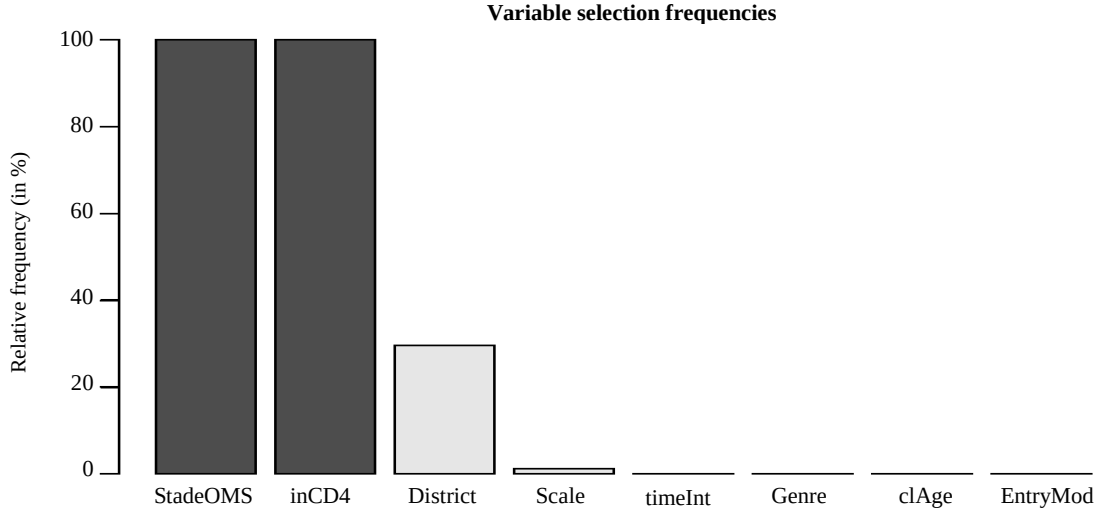

Fig. 5: Stability of variable selection for partioning. Variables StadeOMS, inCD4, colored in dark gray were selected in the original tree. The variables District and Scale were selected by 30% and 1.2% of the bootstrapped trees respectively. They were not selected in the original tree

About 30% of bootstrap trees have selected the variable District for splitting in addition to StadeOMS and inCD4 (Table 3 and Fig. 6). This occurred mostly in the subgroup of patients with CD4 count > 50 cells/μL.

In most cases, the Boulmiougou district forms a first node and Bogodogo and other districts are directed to the second node. As District was not selected by the original tree and its selection frequency is less than 50%, this means that this variable may carry some information that is useful for predicting survival, but it is not among the most important ones. For instance, patients in the Boulmiougou district were followed-up by Médecins Sans Frontières, an NGO from Luxembourg. As a result, health workers involved in the care of HIV infected persons in that district received salaries three times higher than those of other public health centers (Perelman, 2003). Thus, patients may receive a better care compared to patients in the rest of the districts. Finally, very few bootstrap trees (1.2%) have selected the variable Scale for splitting. It need not be retained.

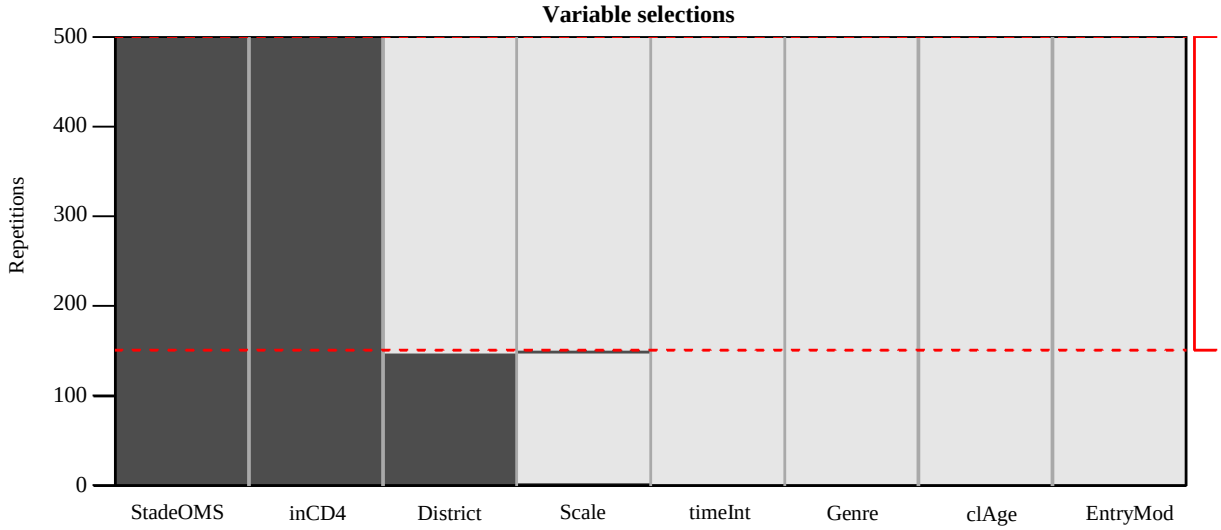

Fig. 6: Frequencies of the different trees built over the repetitions. Dashed horizontal two red lines mark the frequency of the original tree. It is enclosed by a solid vertical red line at the right of the plot

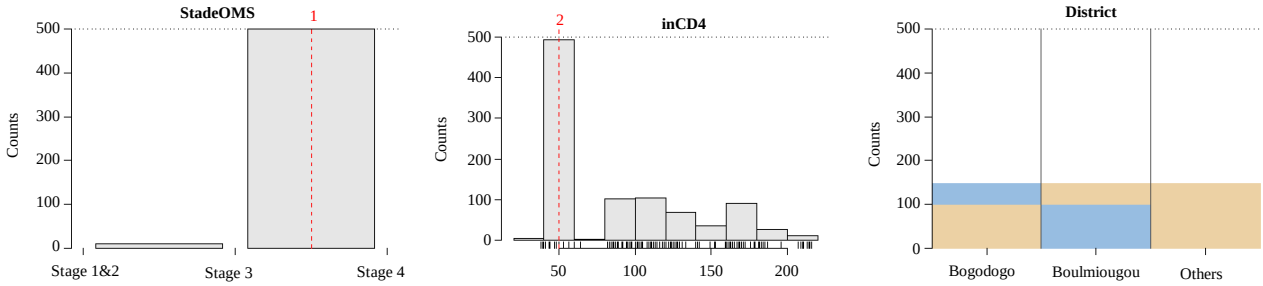

Fig. 7: Stability of the cutpoint selection for partioning. Dashed vertical lines mark the original tree cutpoints. The number above a dashed vertical red line indicates the level at which the split occurred in the tree. For the variable District, categories that belong to the same subgroup are illustrated by the same color

Table 3: Variable selection overview

	Relative frequency (%)	Selected by initial tree	Mean frequency of selections per tree	Selection frequency in initial tree
StadeOMS	100	yes	1	1
inCD4	100	yes	1.9	1
District	30	no	0.3	0
Scale	1.2	no	0.012	0

Concluding Discussion and Remarks

In this study, we proposed a tree-based approach for the analysis of discrete time-to-event data. The method is related to previously proposed methods by (Bou-Hamad *et al*., 2009; Schmid *et al*., 2016). But there are important differences between our method and the two others. First, our method distinguishes two groups of covariates: those used to build a binary tree and those that define a linear model in the nodes of the binary tree; the former are called moderators and the latter have kept the qualifiers of covariates. For example, unlike (Schmid *et al*., 2016) where Time is a splitting covariate (moderator variable in our approach) and each terminal node corresponds to one time interval, our method uses Time as a model covariate and each terminal node corresponds to the whole set of time intervals. When Time alone is included as a covariate, our model is reduced to the Bou-Hamad *et al.* (2009) model.

The second difference is found in the algorithm used to fit the model. An important advantage of our method is that it uses Model-Based algorithm (MOB) (Zeileis *et al*., 2008). The objective is to search for subsets of the data that have the best fits of the hazard model, assuming that the model may not fit perfectly all the dataset. It turns out that each leaf is associated with a fitted model. The algorithm uses the model likelihood function both for parameter estimation and split point search. A benefit of this approach is that the parameter estimates and the corresponding score functions have to be evaluated once in a node. Score functions are then reordered and aggregated into a scalar test statistic each time a parameter instability test is performed. Finally, a Model-Based algorithm is suitable for the identification of subgroups of individuals with similar survival behaviors (Seibold *et al*., 2016).

We have used our method for the analysis of HAART data from Burkina Faso. The model has identified three subgroups of patients with different survival behaviors. The subgroups are determined by the combination of baseline WHO clinical stage and baseline CD4 count. They differ in the shape of the hazard function as well as in the existence and the amount of correlation between the hazard function and at least one predictor variable among Age, Gender and Entry Mode. In each subgroup, the hazard of death is highest in the 1st semester. This early mortality is probably explained by late presentation. Most patients started ART with an advanced HIV infection level. The median baseline CD4 count was 122 (60; 180). As expected, the WHO clinical stage 4 subgroup had the highest within 6-months mortality. Some of the clinical criteria used to assign the disease stage 4 were found to be the main causes of ART patients' deaths in Burkina Faso (Kouanda *et al*., 2011). Patients in other WHO clinical stages that had a low baseline CD4 count (CD4 count < 50 cells/μL) were the second high-risk subgroup. CD4 count < 50 cells/μL was found to be associated with a higher risk of death in other studies (Lawna *et al*., 2008; Kouanda *et al*., 2011). Gender and Age identified as correlated with the hazard function in the subgroup of patients with WHO clinical stage ≤ 3 and baseline CD4 > 50 cells/μL have been reported by other studies as predictors of mortality (De Beaudrap *et al*., 2008; Kouanda *et al*., 2011). Bila and Egrot (2008) reported that the representations of masculinity in Burkina Faso are a factor of men's reluctance to attend health care for persons living with HIV/AIDS (Bila and Egrot, 2008). On the other hand, in each subgroup, categories of patients' hazard profiles differ by the hazard of death estimate in the 1st semester. Patients supported by NGOs or Relatives have lower hazards of death. Social support is important for persons living with HIV/AIDS in the Subsaharan context, where fear of exclusion may lead to poor adherence to treatment (Merten *et al*., 2010). In contrast, transferred patients were found to have the highest hazard of death in the semester 1. The explanation can be the fact that transfer generally occurs in emergency circumstances. The stability analysis showed that a slight instability occurred in the third subgroups of patients defined by WHO clinical stage ≤ 3 and CD4 count > 50 cells/μL. In contrast, the subgroup of patients defined by WHO clinical stage 4 and the one defined by WHO clinical stage ≤ 3 and CD4 count ≤ 50 cells/μL were stable. Therefore the fitted model is fairly stable. On the other hand, we have analysed data for HIV patients on treatment in the thirteen health districts selected for the evaluation of the 5-YEAR Global fund program and in the Boulmiougou district. So our findings should be valid for HIV patients in Burkina Faso and in low-income Subsaharan countries with a similar health system.

Acknowledgment

A large part of the work was carried out at the Laboratoire de Mathématiques et de leurs Applications de Pau (LMAP). We are grateful to the members of LMAP for the facilities and resources they have made available to me. We express our special gratitude to Prof. Marc Artzrouni and to Mrs Marie Henriette Somda for their fine work in the correction of the English text in the manuscript.

Author Contributions

Simon Tiendrébéogo: Analyzed data, wrote analysis tools and wrote the paper.

Blaise Somé: Contributed to write the paper and reviewed the manuscript.

Séni Kouanda: Contributed to write the paper and provide guidance on the clinical interpretation of the findings.

Simplice Dossou-Gbété: Contributed to the data analysis and to the writing of both analysis tools and the paper.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that the other authors have read and approved the manuscript and no ethical issues involved.

References

Berger, M. and M. Schmid, 2017. Semiparametric regression for discrete time-to-event data. Stat. Model.,18: 322-345. DOI: 10.1177/1471082X17748084

Bila, B. and M. Egrot, 2008. Accès au traitement du sida au Burkina Faso: Les hommes vulnérables? Science et Technique, Sciences de la Santé.

Bou-Hamad, I., Denis Larocque, H. Ben-Ameur, L.C. Mâsse and Frank Vitaro *et al*., 2009. Discrete-time survival trees. Can J. Stat., 37: 17-32. DOI: 10.1002/cjs.10007

De Beaudrap, P., J.F. Etard, R. Ecochard, A. Diouf and A.B. Dieng *et al*., 2008. Change over time of mortality predictors after HAART initiation in a Senegalese cohort. Eur. J. Epidemiol., 23: 227-234. DOI: 10.1007/s10654-007-9221-3

Fokkema, M., N. Smits, A. Zeileis, T. Hothorn and H. Kelderman, 2018. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behav. Res. Methods, 50: 2016-2034. DOI: 10.3758/s13428-017-0971-x.

Hastie, T. and R. Tibshirani, 1993. Varying-coefficient models. J. R Stat. Soc, 55: 757-796. DOI: 10.1111/j.2517-6161.1993.tb01939.x

Hornik, K. and A. Zeileis, 2007. Generalized m-fluctuation tests for parameter instability. Inform. Syst., 61: 488-508. DOI: 10.1111/j.1467-9574.2007.00371.x

Hothorn, T. and A. Zeileis, 2016. Partykit: A modular toolkit for recursive partytioning in r. J. Mach. Learn. Res., 16: 3905-3909.

Hothorn, T., K. Hornik and A. Zeileis, 2006. Unbiased recursive partitioning: A conditional inference framework. J. Comput. Graph Stat., 15: 651-674. DOI: 10.1198/106186006X133933

Kouanda, S., 2008. Evaluation des 5 années du fonds mondial: Evaluation globale des districts, [5years evaluation of global fund: District comprehensive assessment]. Technical Report, Institut de Recherche en Sciences de la Santé. Ouagadougou.

Kouanda, S., I.B. Meda, L. Nikiema, S. Tiendrebeogo and B. Doulougou *et al*., 2011. Determinants and causes of mortality in HIV-infected patients receiving antiretroviral therapy in Burkina Faso: A five-year retrospective cohort study. AIDS Care, 24: 478-90. DOI: 10.1080/09540121.2011.630353

Lawna, S.D., A.D. Harries, X. Anglaret, L. Myer and R. Wood, 2008. Early mortality among adults accessing antiretroviral treatment programmes in sub-Saharan Africa. Acquired Immunodeficiency Syndrome, 22: 1-20. DOI: 10.1097/QAD.0b013e32830007cd.

Merten, S., E. Kenter, O. McKenzie, M. Musheke and H. Ntalasha *et al*., 2010. Patient-reported barriers and drivers of adherence to antiretrovirals in sub-Saharan Africa: A meta-ethnography. Tropical Med. Int. Health, 15: 16-33. DOI: 10.1111/j.1365-3156.2010.02510.x

Perelman, B., 2003. Les associations de lutte contre le sida à ouagadougou: Contexte d'émergence, profils, pratiques. Mathesis, Université Paris I.

Philipp, M., A. Zeileis and C. Strobl, 2016. A toolkit for stability assessment of tree-based learners. Proceedings of the 22nd International Conference on Computational Statistics, (CCS' 16), pp: 315-325.

Schmid, M., H. Küchenhoff, A. Hoerauf and G. Tutz, 2016. A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. Stat. Med., 35: 734-751. DOI: 10.1002/sim.6729

Seibold, H., A. Zeileis and T. Hothorn, 2016. Model-based recursive partitioning for subgroup analyses. Int. J. Biostat., 12: 45-63. DOI: 10.1515/ijb-2015-0032

Singer, J.D. and J.B. Willett, 1993. It's about time: Using discrete-time survival analysis to study duration and the timing of events. J. Educ. Behav. Stat., 18: 155-195. DOI: 10.3102/10769986018002155

Zeileis, A., T. Hothorn and K. Hornik, 2008. Model-based recursive partitioning. J. Comput. Graph Stat., 17: 492-514. DOI: 10.1198/106186008X319331

Zelterman, D., C.T. Le and T.A. Louis, 1996. Bootstrap techniques for proportional hazards models with censored observations. Stat. Comput., 6: 191-199. DOI: 10.1007/BF00140864