Applications of Phase Type Survival Trees in HIV Disease Progression Modelling

Marija Gafa
Faculty of Information and Communication Technology
University of Malta
Msida, Malta
marija.gafa.12@um.edu.mt

Lalit Garg
Faculty of Information and Communication Technology
University of Malta
Msida, Malta
lalit.garg@um.edu.mt

Giovanni Masala
Faculty of Economics
University of Cagliari
Cagliari, Italy
gb.masala@unica.it

Sally I. McClean
School of Computing and Information Engineering
University of Ulster
Coleraine, UK
si.mcclean@ulster.ac.uk

***Abstract*— It is important to model progression of a disease to understanding if the patient's condition is improving or getting worse. In the case of HIV disease, the change in the patient's CD4+ T cell count is used to calculate the progression of HIV disease i.e. if the CD4 count goes down it represent the progression of the patient's HIV disease. Due to the lack of an effective cure for HIV disease, it is crucial to monitor the disease progression to managing HIV disease effectively. Therefore, this study is aimed to model HIV disease progression by using phase type survival trees to cluster patients into homogenous groups based on their disease progression to understand the effect of different factors of prognostic significance and their interactions affecting the disease progression. The proposed methods are evaluated using an empirical data of 1,838 HIV-infected patients. The methods developed in this study can also be used for modelling the progression of other chronic conditions or diseases.**

***Keywords*— HIV disease progression, phase type survival trees, disease progression modelling, AIDS**

I. Introduction

Human Immunodeficiency Virus (HIV) is a virus but unlike normal viruses it cannot be cleared out of the body by the immune system, hence once a person is infected with HIV disease he/she will remain with this particular virus for the rest of his/her life. HIV disease attacks the T-cells, also known as the Cluster of Differentiation 4 (CD4) cells, and the CD4+ T cell count is used as the main predictor of disease progression and survival. These cells are a fundamental part of the immune system and HIV disease uses them to make copies of the same virus before it destroys them. Since these cells are crucial to fight infections and diseases, a person infected with HIV disease ends up having a weak immune system and once the immune system becomes deficient, hence the name, it cannot protect the patient against other viruses. Furthermore, when a certain amount of CD4 cells are destroyed, HIV disease becomes Acquired Immunodeficiency Syndrome (AIDS), which is the final stage of HIV disease. However, it should be noted that not every person who has HIV disease ends up having AIDS since antiretroviral treatment (ART) increases the number of CD4 cells, the level of HIV virus in the body is kept low and as a result AIDS is prevented. Nowadays a person can have a normal life expectancy if he/she is diagnosed with HIV disease and treated before the disease is far advanced. However, poor monitoring of disease progression and HIV disease drug resistance, i.e. the inability of the antiretroviral drug to decrease the viral reproduction rate adequately, are the main causes of HIV infected patients' fatalities. Therefore, although HIV disease can be managed using ART, monitoring closely the disease progression is crucial.

In this research work, HIV infected patients are clustered into groups using phase type survival trees (PTSTs). This clustering is done with respect to the total duration that a patient survives after being diagnosed with HIV disease and also with respect to the length of stay (LOS) of a patient in a particular HIV state. Partitioning is based on covariates representing some of the patients' characteristics such as age, gender, therapy, censoring, start state, end state, initial state and final state. This is done to obtain a covariate which is the best to group the patients, i.e. which covariate has the maximum effect on the survival duration or on the LOS in an HIV state.

II. AIMS AND OBJECTIVE

Since currently no effective cure for HIV disease exists, the aim behind this research work is to model the progression of HIV disease as to identify the factors or covariates which contribute to the progression of the disease using phase type distribution (PHD). This is done by using a real database of 1,838 HIV infected patients which were enrolled in the Italian public structures from January 1996 to January 2008. The outcome should aid to manage HIV disease and its treatment in a more effective way and can help us to understand better the effects that different covariates have on HIV disease progression while the methods developed through this dissertation can be useful for modelling disease progression in patients who suffer from other diseases.

III. BACKGROUND AND RELATED WORK

A. *HIV Disease Progression*

In a person who has a healthy immune system, i.e. a person who is HIV negative, the CD4 counts are between 500 and 1,500 cells/mm^3, i.e. cells per cubic millimeter of blood, and in the case of an HIV positive person the CD4 count decreases as the HIV disease continues to advance, hence, HIV disease is

978-1-5386-3893-4/17/$31.00 ©2017 IEEE

termed as progressing if the CD4 count goes down. In the case of an HIV positive person, ART is used to prevent the HIV virus from multiplying and from destroying the patient's immune system which consequently slows HIV disease progression. This helps in keeping the patient's body healthy to fight off life-threatening infections and prevents HIV disease from progressing to AIDS.

The CD4+ T cell count is used as the major predictor of HIV disease progression and survival and in fact the US Department of Health and Human Services (DHHS) ART treatment guidelines suggest that initiation of treatment should be based on this count [1].

HIV disease has several stages, where each stage is determined solely by the patient's absolute peripheral blood CD4 count, and these stages indicate and describe the disease severity and the disease progression. The World Health Organisation (WHO) published a general disease staging system for these HIV disease stages. The staging system is however different for children, adolescents and adults. The four prognostic stages are [2, 3]:

1. Asymptomatic disease where CD4 levels >500/mm^3.

2. Mild/ minor disease where CD4 levels 350-499/mm^3.

3. Moderate/advanced disease where CD4 levels 200-349/mm^3.

4. AIDS/severe disease where CD4 levels <200/mm^3.

It should be noted that a patient's immunological status may in some cases progress sequentially through the stages, however, it might also jump from a particular stage to any other stage. This disease progression from one stage to another depends on many factors and hence people infected may progress through the stages of HIV disease at different rates [1, 4].

It should also be noted that the total time from the beginning of the infection till the infection develops to AIDS is known as the incubation period while the time from when HIV is diagnosed till death is known as HIV survival time. To find tests that are useful for prognosis and treatment decisions in HIV disease, and to establish what factors and covariates increase or decrease the rate of HIV disease progression, the survival period and the distribution of possible lengths of the incubation period needs to be characterised [5].

B. Phase Type Distribution(PHD)

In this research work, PHD will be used to model HIV disease progression and this can be done by modelling and fitting PHD to its data. "*A phase type distribution(PHD) is defined as the distribution of the lifetime X, i.e. the time to enter an absorbing state from the set of transient states ST of an absorbing continuous time Markov process*$\{X(t)\}t^{\infty}\geq 0$" [6].

PHDs are an extension of the exponential distributions, in fact they show many of their advantageous properties, for example PHD, like exponential distributions, have the memoryless property, i.e. they are defined only on the nonnegative real numbers. PHD is frequently used in a wide range of application areas to model Markov stochastic process and they can model realistically the process of a patient's journey through several stages as a finite state continuous time Markov chain [7, 8].

The analysis of healthcare systems is an important application area where PHD can be applied since in this area PHD is used to describe the time that patients stay in hospitals or to describe infection models, where in the latter PHD models the duration of different phases of an infection [6, 9].

However, PHD models are very flexible and as a consequence huge effort to find the parameters is needed so that the resulting model approximates closely the required or observed behavior.

C. Coxian Phase Type Distribution(C-PHD)

The best fit is known to be provided by the general PHD [8], however, parameter estimation is difficult in this type of distribution. The Coxian phase type distribution (C-PHD) is a type of PHD that provides a suitable fit and it does not present the problems of the general PHD. This is because C-PHD has the minimum number of estimable parameters and it also has the advantage that it provides a simple interpretation of fit. C-PHD is a special type of PHD where entry can only occur to the first state, i.e. a patient can only enter the system in the first state and only sequential transitions can take place, as shown in Figure 1 which was obtained from [9], with a transition rate λ_k from any state k to the next state k+1. A transition from any state k to the absorbing state n+1 is also possible with a transition rate μ_k where the absorbing state represents the death of a person [10].

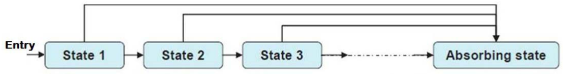


Fig. 1. Schematic representation of C-PHD

The time that a patient survives after being diagnosed with HIV disease has probability density function

$$f(t) = p(exp(Qt))q; \tag{1}$$

where p, which represents the initial state probability distribution, is defined as

$$p = (1\ 0\ 0\ .\ .\ .\ 0\ 0); \tag{2}$$

the transition matrix Q is defined as

$$\begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & \cdots & 0 \\ 0 & -(\lambda_2+\mu_2) & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & -\mu_n \end{pmatrix} \tag{3}$$

and q, which represents absorption probabilities, is defined as

$$q = (\mu_1\ \mu_2 \ldots\ \mu_{n-1}\ \mu_n)^T \tag{4}$$

The likelihood function is used to choose the best fit for the distribution to the data and the log-likelihood function, which is more convenient to work with, is defined as

$$L = \sum_{i=1}^{N} (\log(p \exp\{Q t_i\} q)) \quad (5)$$

where N is the sample size [10].

D. Phase Type Survival Trees(PTSTs)

Survival trees are trees used in survival analysis and they are a type of classification and regression trees. Survival tree based analysis is a robust method of splitting the data into subgroups, based on covariates such as age, for prognostication i.e. for establishing the effects of different input covariates and their effects on output measures such as disease progression or a patient's survival duration.

PTSTs are a special type of survival trees where each node in the tree is described separately by PHD. A PTST is constructed by recursively splitting nodes into subgroups by one of the covariates based on some splitting criteria [11].

Various splitting criteria to construct a PTST exist, for example splitting criteria based on the maximum likelihood ratio or the Bayesian information criterion (BIC) or the weighted information criterion (WIC) based splitting criteria [10, 11, 12].

In [9], the authors use PTSTs to evaluate the effect that the HIV diagnosis (HD) stage has, using the akaike information criterion (AIC). From this study, it was analysed that after clustering the data with respect to the HD stage there was a significant improvement in AIC and hence the HD stage affects the LOS that a patient spends in an HIV infection (HI) stage. This approach is then extended by first clustering the data based on the HI stage and then each cluster is further sub-clustered based on the HD stage. This approach also enhanced the analysis that the patient's HD stage affects the LOS in a particular HI stage and it was also observed that patients who were diagnosed in stage 4 stay in HI stage 4 for a longer period of time than any of the other stages, and similarly for the other HD stages.

Furthermore, in another study [11], the authors proposed a method for clustering the patients into homogeneous groups using PTSTs with respect to their LOS. However, in this paper the splitting criteria used is based on improvement of log-likelihood functions and not on the WIC splitting criteria as in the case of this dissertation. In their other paper [10], the authors illustrate how the approach used in [11] can be used to quantify the significance of different covariates and their interaction in forecasting the patient's LOS in the hospital. This paper then describes how the PTST method can be extended and used to examine the relationship between the LOS in hospital and destination on discharge. In this approach, the BIC is used and this paper illustrates that this extended PTST approach can be used effectively to analyse the relationship between destination at discharge and LOS. This helps for estimating bed requirements and the cost of care while our approach helps to examine the disease progression.

IV. SPECIFICATION AND DESIGN

In this research work, patients are clustered into groups using PTSTs. Firstly, the clustering is done with respect to the total duration that a patient survives after being diagnosed with HIV disease and secondly, the clustering is done with respect to a patient's LOS in a particular HIV state. This was done using a real dataset consisting of 1,838 HIV infected patients with a total of 9824 examination visits which were enrolled in the Italian public structures from January 1996 to January 2008. No information that identified individual patients was provided.

Partitioning is based on covariates representing some of the patients' characteristics such as age, gender, therapy, censoring, start state, end state, initial state and final state. This is done to obtain a covariate which is the best to group the patients, i.e. which covariate has the maximum effect on a patient's survival duration and on a patient's LOS in an HIV state.

A. General Procedure for Clustering Patients into Groups Based on the WIC Splitting Criteria

A In this research work, PTSTs were constructed using the WIC based splitting criteria to cluster and identify the effects of various covariates and their interaction in the prediction of HIV disease progression. WIC is defined as follows:

$$WIC(d) = -2(Loglikelihood) + d + (\frac{d(((\log(N) - 1)\log(N))(N - (d+1))^2 + 2N(N + (d+1)))}{(2N + (\log(N)(N - (d+1))))(N - (d+1))}) \quad (6)$$

where N is the number of patients and d = 2*(k-1) where k is the number of phases [13].

This particular splitting criteria is used throughout this research work since it was observed in [14] that it produces the best results. Also in another study [15], it was found that WIC performs at least as well as, and in some cases even outperforms, other splitting criteria both with small sample and large sample sizes.

It should be noted that a variant of the freely available downloadable package EMpht [16] is used to estimate parameters for the C-PHD fit to the duration and LOS. This package uses the expectation-maximization (EM) algorithm to implement maximum likelihood parameter estimation.

Since the age covariate is a continuous covariate it should be divided into groups as to make it a categorical covariate. A PTST is then constructed by recursively splitting nodes into subgroups by one of the covariates. This is done by fitting each group separately to C-PHD and then the WIC value is used as a splitting criteria for the PTST. Since C-PHD is used, the covariate which provides the minimum total WIC value is chosen for further partitioning where the total WIC value for a particular covariate a is the sum of the minimum WIC values, which is given by a particular phase, for each subgroup of a particular covariate. The minimum WIC value chosen must be

$\geqslant -2*$loglikelihood for the WIC value to be correct and if otherwise that WIC value is not considered and the next smallest WIC value from another phase is taken. If no WIC value is correct, the WIC value for that group is considered as null for the addition.

Once a covariate a which provides the minimum WIC value, hence the maximum improvement in WIC, is obtained, this covariate a splits the dataset into a number of subgroups and then each subgroup is separately fitted to the C-PHD. One covariate is applied at each node and the total WIC value for partitioning with that covariate is recorded. This is repeated with other covariates and then the WIC value is compared with the WIC value of the node before partitioning. Hence a covariate which provides maximum significant improvement, by minimizing the WIC, is selected by exploring all possible splits each time. If at a particular node there is no split which provides significant improvement in WIC, that node is considered a terminal node and hence the process above should be repeated until no node provided a split which provides a significant improvement in WIC.

B. Modelling a Patient's Survival Duration after HIV Disease Diagnoses Headings

The survival duration of each patient, i.e. the total time from when the patient was diagnosed with HIV disease till death or till termination of data collection was modelled. Partitioning is based on the four covariates age, gender, initial state and therapy. For the continuous covariate age, cut-points were used to divide patients into groups as to change the continuous covariate age into a categorical one.

The age covariate was divided into 3 almost equal subgroups using 3 cut-points as shown below:

First Group: 15.75 to 25.99 years, 598 patients

Second Group: 26 to 33.99 years, 755 patients

Third Group: 34.03 to 75.36 years, 485 patients.

A PTST, shown in Figure 2 Section IV, was created to model a patient's survival duration using the procedure explained in Section IV.A to cluster the patients into meaningful groups. This was also repeated to model a patient's LOS in an HIV state.

V. RESULTS

Figure 2 is the schematic diagram of the PTST that was constructed using the WIC based splitting criteria and it clusters the data into 8 clusters (i.e. terminal nodes). This PTST implies that the whole data was first divided by the therapy covariate since it was found to have a high prognostication significance on the patient's survival duration, then it shows which other covariates have a prognostic significance on the survival duration when excluding therapy, and hence it further considers patients who did not use ART (node 2) and patients who did use ART (node 3). The other nodes have the same interpretation and illustrate what other covariates can affect the patient's survival duration. This PTST can help us determine factors which can affect the survival duration of a newly diagnosed HIV infected patient.

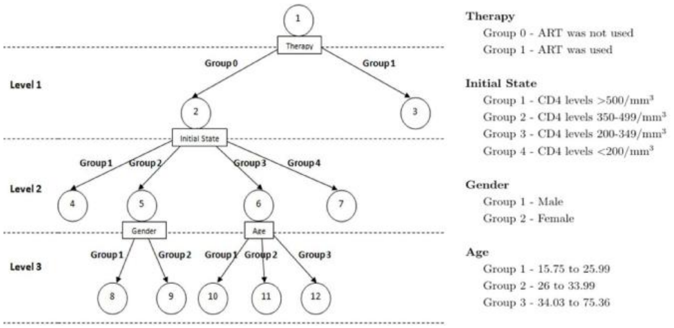


Fig. 2. PTST modelling a patient's survival duration

Table 1 presents the nodes of the tree together with all the possible splits of these nodes. The splits selected to create child nodes are shown as bold faced. For instance nodes 8 and 9 are nodes created by splitting node 5 by the gender covariate. The gain in WIC is the difference between the WIC value of the parent node before clustering and the total WIC value of the leaf nodes. Table 1 also lists the number of patients in each group together with the mean duration. This table provides 8 prognostic groups from the survival data used where each group follows a separate patient pathway and one can analyse the relationship between age, gender, initial state, therapy and total duration by analysing the results in this table.

At level 1 it can be observed that the most significant split is given by the covariate therapy (total WIC 10511.97) which means that there was most significance difference among patients who did not use ART (Group 0) and those who did (Group 1). Patients who did not use ART have a shorter survival duration (mean duration 1640.57) while patients who used ART are less likely to have shorter duration (mean duration 1720.66) showing that the use of ART increases the patient's survival duration. The second best splitter at level 1 is the covariate initial state (total WIC 10534.13). Patients who were diagnosed in state 4, i.e. initial state = 4, were most likely to have a short survival duration (mean duration 1104.25) while patients who were diagnosed in state 2, i.e. initial state = 2, were less likely to have a short survival duration (mean duration 1978.97). The other two covariates age and gender did not provide any gain in WIC at this level.

At level 2, the covariate initial state provided the most significant split for node 2 (total WIC 3945.44). In this case it can be observed again that patients who were diagnosed in state 4, i.e. initial state = 4, were most likely to have a short survival duration (mean duration 627.54) while patients who were diagnosed in state 2, i.e. initial state = 2, were less likely to have a short survival duration (mean duration 1946.68). No covariate provided an improvement in WIC for node 3 and hence node 3 is a leaf node in the PTST.

At level 3, no covariate provided an improvement in WIC for nodes 4 and nodes 7 and hence these nodes are also terminal nodes. The covariate gender provided the most significant splits for node 5 (total WIC 681.76). In this case it can be observed that females were most likely to have a short survival duration (mean duration 1677.27) while males were less likely to have a short survival duration (mean duration 2052.39).

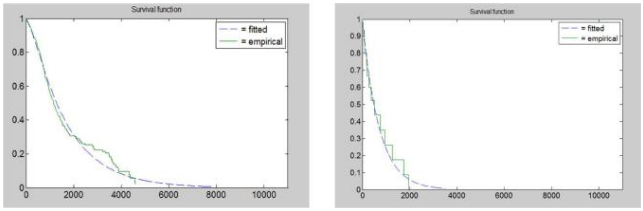
TABLE I. TREE CONSTRUCTION TABLE MODELLING A PATIENT'S SURVIVAL DURATION

Node	Covariate	Covariate Value	No. of patients	Mean Duration	WIC	Log likelihood	Total WIC	Gain in WIC
All	**Complete dataset**	**Root node**	**1838**	**1868.77**	**10536.12**	**-5264.88**	**10536.12**	**-**
1	Age	1	598	2031.60	3486.44	-1740.55	10540.54	-4.42
		2	755	1926.39	3944.72	-1969.58		
		3	485	1619.33	3109.37	-1552.11		
	Gender	1	1288	1829.25	8041.04	-4017.50	10540.20	-4.08
		2	550	1997.41	2499.16	-1246.94		
	Initial State	1	1030	1964.99	5842.44	-2918.30	10534.13	1.99
		2	425	1978.97	1980.82	-987.89		
		3	274	1866.02	1762.16	-878.75		
		4	109	1104.25	948.70	-472.41		
	Therapy	**0**	**504**	**1640.57**	**3954.51**	**-1974.66**	**10511.97**	**24.15**
		1	**1334**	**1720.66**	**6557.46**	**-3269.63**		
2	Age	1	173	1877.23	1319.05	-657.39	3956.29	-1.78
		2	212	1698.15	1641.28	-818.42		
		3	119	1250.26	995.95	-496.00		
	Gender	1	368	1631.61	2960.78	-1477.93	3958.28	-3.77
		2	136	1667.30	997.51	-496.72		
	Initial State	**1**	**322**	**1780.23**	**2414.40**	**-1204.80**	**3945.43**	**9.08**
		2	**133**	**1946.68**	**682.09**	**-335.17**		
		3	**45**	**1502.87**	**518.74**	**-257.77**		
		4	**24**	**627.54**	**330.21**	**-163.72**		
3	Age	1	425	1745.23	2164.58	-1074.72	6574.71	-17.24
		2	543	1690.63	2298.20	-1141.21		
		3	366	1800.89	2111.93	-1053.51		
	Gender	1	920	1733.21	5073.26	-2528.03	6568.60	-11.14
		2	414	1574.16	1495.34	-740.14		
	Initial State	1	708	1793.76	3423.68	-1703.60	6580.16	-22.70
		2	312	1760.67	1298.74	-642.21		
		3	229	2022.37	1244.64	-620.07		
		4	85	1387.70	613.11	-304.71		
4	Age	1	129	1805.10	989.84	-492.91	2421.06	-6.66
		2	134	1768.08	1021.38	-508.66		
		3	59	1750.50	409.85	-203.22		
	Gender	1	230	1788.37	1753.26	-874.37	2417.85	-3.45
		2	92	1758.74	664.59	-330.42		
5	Age	1	30	2553.00	179.82	-88.45	687.13	-5.04
		2	50	1643.00	289.02	-142.87		
		3	33	1452.69	218.29	-107.66		
	Gender	**1**	**83**	**2052.39**	**493.49**	**-241.24**	**681.76**	**0.33**
		2	**30**	**1677.27**	**188.26**	**-92.67**		
6	**Age**	**1**	**11**	**1881.99**	**121.98**	**-59.78**	**514.96**	**3.78**
		2	**19**	**2207.45**	**194.03**	**-95.70**		
		3	**15**	**702.54**	**198.95**	**-98.21**		
	Gender	1	36	1424.69	432.65	-214.80	520.59	-1.86
		2	9	1909.39	87.94	-42.77		
7	Age	1	3	573.50	34.65	-14.70	337.24	-7.03
		2	9	713.66	138.67	-68.13		
		3	12	566.91	163.93	-80.74		
	Gender	1	19	667.94	272.79	-135.08	332.36	-2.15
		2	5	445.75	59.57	-28.40		
8	Age	1	21	2411.12	143.30	-70.30	500.26	-6.77
		2	37	2635.54	175.38	-83.06		
		3	25	-89.39	181.58	-89.39		
9	Age	1	9	3120.50	38.58	-18.09	189.64	-1.38
		2	13	1001.85	113.20	-55.37		
		3	8	2598.00	37.86	-17.72		
10	Gender	1	8	1603.33	102.97	-50.28	126.56	-4.58
		2	3	3553.99	23.59	-9.18		
11	Gender	1	15	2072.66	157.98	-77.73	197.08	-3.05
		2	4	2813.99	39.10	-17.88		
12	Gender	1	13	797.09	171.45	-84.49	-	-
		2	2	-	-	-		

No covariates provided an improvement in the WIC value for nodes 8,9,10,11 and 12 and hence they are labelled as terminal nodes as well. It should be noted that a '-' in Table 1 implies that no WIC value was $\geqslant -2*$loglikelihood.

This shows that the covariate therapy has a prognostic significance since splits based on this covariate provide significant improvement in WIC. Therapy is followed by initial state for those patients who did not use ART while no covariate provides any improvement in WIC for those who used ART and hence this cluster of patients is not further divided. Also no covariates provided an improvement for initial state groups 1 and 4 while group 3 and group 4 are furthered partitioned by gender and age respectively, however, these two covariates provide a lower prognostic significance compared to the therapy covariate.

The survival function was then plotted for each terminal node to show the quality of fit of each cluster of patients and the plots obtained verified that the model created represents the empirical data well (See Figure 3). The survival functions for node 3 and node 12 are shown in Figures 3(a) and 3(b) respectively where the x-axis represents the number of days that a patient survives after being diagnosed with HIV disease while the y-axis represents the probability of survival. Therefore, from such plots we can estimate the probability that a newly diagnosed HIV patient, which belongs to a specific clustered group, has of survival after a certain amount of days of being diagnosed with HIV.



(a) Node 3 (1334 patients) (b) Node 12 (15 patients)

Fig. 3. Survival Function for (a) Node 3, (b) Node 12

TABLE II. EMPIRICAL AND FITTED DATA RESULTS FOR NODE 3 - PATIENTS WHO USED ART

No. of days after HIV diagnosis	*Empirical - Prob. Of survival (%)*	*Fitted – Prob. Of survival (%)*
0	100	100
988	57	60
1853	30	34
2877	22	17
3895	10	9
4580	4	6

From Figures 3(a), it can be observed that on day 1 of diagnosis there is 100% chance of survival while as the number of days increase, the survival function decreases. This implies that the probability of survival decreases as the number of days increase.

The values of the x-axis and the corresponding y-axis values for Figure 3(a) were recorded, i.e. for patients who used ART, for both the empirical data and for the fitted data. The results are shown in Table 2.

The results obtained verify that the model obtained gives a good prediction for the cluster of patients who used ART since for instance after 988 days a patient has a probability of survival of 57% while the model created by fitting the data gives a probability of survival of 60%. It can also be noted that the model created predicts that after approximately 8000 days (approximately 22 to 23 years) a patient who uses ART (excluding any other covariate) has 0% of survival. On the other hand, from Figure 3(b) it can be observed that a 34 to 75 years old patient who was diagnosed with HIV in state 3 and does not use ART (node 12) has 0% of survival after 3000 days (approximately 8 to 9 years).

It should be noted that another PTST modelling the patient's LOS in an HIV state was created and it was obtained that the factor end state, which represents the state in which the patient was during a particular examination visit, has a high prognostic significance when modelling the LOS in an HIV state. However, in this case the data is censored as the CD4 count is measured at each examination visit and the HIV state is diagnosed based on this count but the patient might have been in this state even before the examination visit. As a result one cannot know the exact time that a patient entered or left a particular HIV state and therefore the data is censored, except when a patient dies

VI. EVALUATION

The aim of this research was to model HIV disease progression and to identify the factors or covariates which contribute to the progression of this disease and as shown in Section V, the covariate therapy has a high prognostic significance followed by initial state while age and gender provide a lower prognostic significance when modelling a patient's survival duration. In this section, 3-fold cross-validation is used for evaluating the model which models a patient's survival duration. Therefore, the original dataset used in this research work was divided into three subsamples and a PTST was created from $\frac{2}{3}$ of the data and then the remaining $\frac{1}{3}$ of the data was used to test the fit of the model. This process was repeated three times and hence three PTSTs were obtained.

A. PTSTs for the Subsamples

The PTSTs created for subsamples 1 and 2 are identical and hence one PTST, Figure 4, is provided for both while the PTST for subsample 3 is illustrated in Figure 5.

All three subsamples give a slightly different PTST from the original PTST , however, these three PTSTs still verify that the covariate therapy has the most prognostic significance on the patient's survival duration, since splits based on this covariate provided significant improvement in WIC even for $\frac{2}{3}$ of the data, followed by the initial state covariate for those patients who did not use ART, as in the original PTST, while no other covariates has any other significance for those patients who used ART for these three PTSTs. Also, no covariate provided an improvement for initial state groups 1 and 4, as in the original PTST.

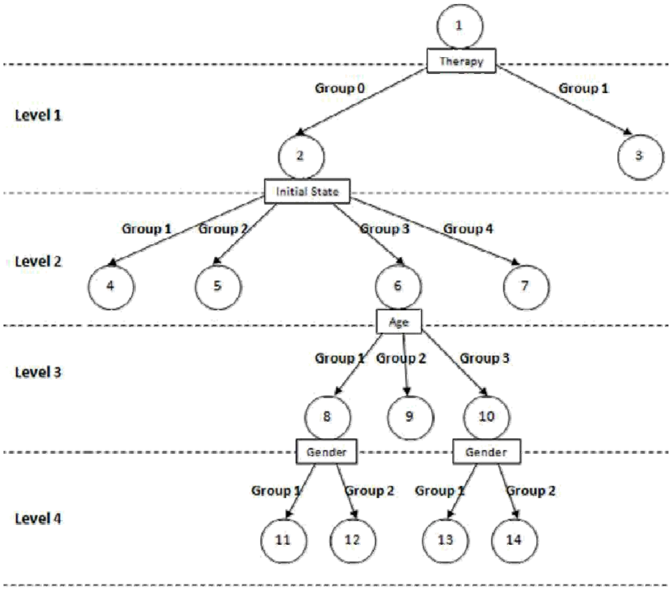


Fig. 4. PTST for subsamples 1 and 2

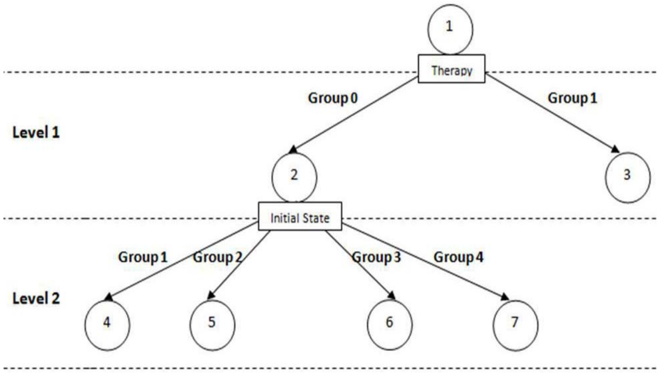


Fig. 5. PTST for subsample 3

The only difference that these PTSTs have from the original PTST, Figure 2 Section V, is shown in the bottom levels of these PTSTs since for subsample 1 and subsample 2, initial state groups 3 and 4 are not further partitioned and hence age and gender provide no further partitioning for these two subsamples. However, for subsample 3, initial state group 3 is furthered partitioned by the age covariate and then age group 1 and age group 3 are furthered partitioned by the gender covariate, however, this shows that these two covariates still provide a low prognostic significance when modelling a patient's survival duration.

B. *Comparing Empirical to Fitted Data*

The distribution plots for the C-PHD fitting of the data, represented by the terminal nodes of the PTSTs, from $\frac{2}{3}$ of the data compared with the corresponding $\frac{1}{3}$ of the data for each node of each PTST were plotted.

This was done to compare the empirical (from $\frac{1}{3}$ of the data) and fitted data (from $\frac{2}{3}$ of the data). the plots obtained illustrate that the model created, based on a patient's survival after being diagnosed with HIV disease, provides a good fit and minimum overfit the data. A slight difference between the empirical and fitted data could be observed in some cases, however, it was due to the fact that the empirical data consists of a small number of patients for those nodes. Therefore, the model created in Section V can also represent data which was not used to estimate its parameters. As a result, it is shown that PTST based analysis can be used to identify predictors of a patient's survival duration.

VII. CONCLUSIONS AND FUTURE WORK

The motivation behind this research work was to model the progression of HIV disease and this was done by creating and evaluating the PTSTs together with illustrating how these PTSTs can be constructed and used to cluster and quantify the significance of different covariates which contribute to the progression of HIV disease. The PTSTs created cluster the patients into prognostically significant groups and are effectively revealing the interrelationship between the covariates and the patient's survival. As a result, when considering the total duration that a patient survives after being diagnosed with HIV disease, the covariate therapy was found to have prognostic significance followed by initial state showing their prognostic significance to the progression of HIV disease while the covariates age and gender did not show any prognostic significance, while when considering the patient's LOS in an HIV state, the covariate end state provided a prognostic significance to the disease progression. This illustrates that PTST based analysis is a practical method for establishing the relationship between input covariates and outcome measures and for clustering available survival data into clinically meaningful patient groups. As a result it is shown that PTST based analysis can be used to identify predictors of a patient's survival duration and to estimate the survival duration of a patient based on his/her characteristics that are available at the time of HIV disease diagnosis.

The outcome of this research is practical since currently no effective cure for HIV disease exists and the results obtained can help us to understand better the effects that different covariates have on HIV disease progression. Consequently, this can help in better planning and care management of the HIV infected patients while it might also help in understanding the effect of ART on the disease progression. Such factors could also help in finding ways to slow disease progression and help in therapeutic monitoring decisions.

As future work, the approach used in this research work can be repeated using different cut-points for the age covariate and an automated algorithm that can be used to decide the optimum cut-points can be developed. The PTST analysis can also be extended by considering other covariates, for instance coinfection with other diseases such as hepatitis B or hepatitis C. Additionally, a similar approach using PTST based analysis can be applied for modelling disease progression in patients who have other infectious diseases such as diabetes and cancer

REFERENCES

[1] S. E. Langford, J. Ananworanich, and D. A. Cooper. Predictors of disease progression in hiv infection: a review. AIDS Research and Therapy, 4, 2007.

[2] W. H. Organization. Interim who clinical staging of hiv/aids and hiv/aids case definitions for surveillance. AIDS Research and Therapy, 2005.

[3] AIDS.gov. Stages of hiv infection, December 2013.

[4] G. Pantaleo, S. Menzo, M. Vaccarezza, C. Graziosi, O. J. Cohen, J. F. Demarest, D. Montefiori, J. M. Orenstein, C. Fox, L. K. Schrager, J. B. Margolick, S. Buchbinder, J. V. Giorgi, and A. S. Fauci. Studies in subjects with long-term non progressive human immunodeficiency virus infection. The New England Journal of Medicine, 332(4):209–216, Jan 1995.

[5] D. H. Osmond. Epidemiology of disease progression in hiv. May 1998.

[6] P. Buchholz, J. Kriege, and I. Felko. Introduction and phase-type distributions. In Input Modeling with Phase-Type Distributions and Markov Models, chapter 2, pages 1–27. Springer International Publishing, 2014.

[7] L. Garg, S. McClean, B. J. Meenan, and P. Millard. Phase-type survival trees and mixed distribution survival trees for clustering patients' hospital length of stay. Informatica, 22(1):57–72, 2011.

[8] M. Fackrell. Modelling healthcare systems with phase-type distributions. Health Care Management Science, 12(1):11–26, 2009.

[9] L. Garg, G. Masala, S. I. McClean, M. Micocci, and G. Cannas. Using phase type distributions for modelling hiv disease progression. In 25th International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2012.

[10] L. Garg, S. McClean, M. Barton, B. Meenan, and K. Fullerton. An extended phase type survival tree for patient pathway prognostication. In IEEE Workshop on Health Care Management (WHCM). IEEE, 2010.

[11] L. Garg, S. McClean, B. Meenan, and P. Millard. A phase type survival tree model for clustering patients' hospital length of stay. In The XIII International Conference, Applied Stochastic Models and Data Analysis (ASMDA), 2009.

[12] L. Garg, S. McClean, M. Barton, B. Meenan, and K. Fullerton. Forecasting hospital bed requirements and cost of care using phase type survival trees. In Intelligent Systems (IS), 5th IEEE International Conference. IEEE, 2010.

[13] L. Garg, S. I. McClean, M. Barton, B. J. Meenan, and K. Fullerton. Intelligent patient management and resource planning for complex, heterogeneous, and stochastic healthcare systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 42(6), 2012.

[14] L. Garg. Unified modelling for care of the elderly. PhD thesis, University of Ulster, UK, 2010.

[15] P. Chen, T.-J. Wu, and J. Yang. A comparative study of model selection criteria for the number of signals. IET Radar Sonar Navigation, 2(3):180–188, 2008.

[16] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distributions via the em algorithm. Scandinavian Journal of Statistics, 23(4):419–441, December 1996.