INFORMATICA, 2011, Vol. 22, No. 1, 57–72

© 2011 *Vilnius University*

Phase-Type Survival Trees and Mixed Distribution Survival Trees for Clustering Patients' Hospital Length of Stay

Lalit GARG[1], Sally McCLEAN[1], Brian J. MEENAN[1], Peter MILLARD[2]
[1] *University of Ulster, Coleraine*
Co. Londonderry, BT52 1SA, UK
[2] *St. George's Hospital Medical School*
12 Cornwall Road, Cheam, Sutton, Surrey, SM2 6DR, UK
e-mail: {l.garg, si.mcclean, bj.meenan}@ulster.ac.uk, phmillard@tiscali.co.uk

Received: October 2009; accepted: January 2011

Abstract. Clinical investigators, health professionals and managers are often interested in developing criteria for clustering patients into clinically meaningful groups according to their expected length of stay. In this paper, we propose two novel types of survival trees; phase-type survival trees and mixed distribution survival trees, which extend previous work on exponential survival trees. The trees are used to cluster the patients with respect to length of stay where partitioning is based on covariates such as gender, age at the time of admission and primary diagnosis code. Likelihood ratio tests are used to determine optimal partitions. The approach is illustrated using nationwide data available from the English Hospital Episode Statistics (HES) database on stroke-related patients, aged 65 years and over, who were discharged from English hospitals over a 1-year period.

Keywords: decision support, clinical databases, phases of care, estimating group service time.

1. Introduction

Decision trees in survival analysis are popularly known as survival trees and are a type of classification and regression tree (Breiman *et al.*, 1984; Davis and Anderson, 1989; Šutienė *et al.*, 2010; Žilinskas and Žilinskas, 2010). Survival trees are special types of mixture distributions where the dataset is partitioned into subgroups based on the values of covariates and each subgroup is approximated by different set of components. Survival tree based analysis is a powerful non-parametric method of clustering survival data for prognostication, i.e., to determine importance and effect of various covariates (such as patients' characteristics) and their interrelation with patients' survival, treatment outcome, disease risk, disease progress or hospital length of stay (Davis and Anderson, 1989; Gao *et al.*, 2004). Zhang and Singer (1999) provide a review of the literature of survival trees and their applications. More recently Clarke and West (2007) used Bayesian Weibull survival tree models for analysis of clinico-genomic data and Fan *et al.* (2009) used multivariate exponential survival tree based analysis for tooth prognosis. However most of the

survival tree based models suffer from a limited ability to realistically model the process dynamics of a care system and are not exactly understood by medical practitioners.

The novel contribution of this paper is to propose two new techniques based on survival trees for clustering patients into clinically meaningful patient groups with respect to their hospital length of stay, where partitioning is based on covariates such as gender, age at the time of admission and disease diagnosed. The first technique is based on phase-type survival trees where each node is modelled by a Coxian phase type distribution. The proposed phase type survival trees can effectively be used for clustering survival data into groups of patients following homogeneous patient pathways. They provide understanding of heterogeneity of patient pathways stratified by covariates representing patient characteristics such as age, gender and diagnosis. The second technique is based on mixture distribution survival trees where different nodes in the tree can be modelled by distinct types of mixture distributions. The mixture distribution survival tree provides better improvement in the likelihood function and in the within node homogeneity. We have used splitting criteria based on improvement of log-likelihood functions. Phase type distributions (PHDs) and Gaussian mixture distributions (GMDs) are among popular choices for modelling hospital length of stay (Fackrell, 2009; Marshall, 2007; El-darzi *et al.*, 2009; Abbi *et al.*, 2008). One key advantage of these new types of survival trees is their ability to accurately model patient pathways followed by different groups of patients as finite state continuous time Markov chain facilitating easy analysis of models and better explainability to healthcare professionals than other survival tree models.

In this paper, first we illustrate how phase type survival trees and Gaussian mixture survival trees can be constructed and used for clustering length of stay data. Then we extend this approach by introducing more flexibility in the tree construction where different nodes in the tree can be approximated by distinct types of mixture distributions. This way, more improvement in the likelihood function can be achieved, which results in the improved within node homogeneity.

The next section provides the background on phase type distributions and Gaussian mixture distribution. The following section illustrates how phase type survival trees and Gaussian mixture survival trees can be constructed. Running examples of phase type survival trees and Gaussian mixture survival trees are illustrated using nationwide data available from the English Hospital Episode Statistics (HES) database on stroke-related patients, aged 65 years and over, who were discharged from English hospitals over a 1-year period. The paper then describes how we can construct and use a mixed distribution survival tree for clustering length of stay data and the method is illustrated using the same HES database as a running example.

2. Background

Gaussian mixture distributions (GMD) and phase type distributions (PHDs) are among the popular choices for approximating patients' hospital length of stay and for determining clinically meaningful patient groups from a given dataset of patients' length of stay.

PHDs' components (or clusters) follow generalized Erlang type distributions while Gaussian mixture distributions assume that their components are multivariate normal (Gaussian) distributed. PHDs are a popular choice for modelling Markov stochastic process in a wide range of application areas (Fackrell, 2009; Marshall, 2007). GMDs are well known and have been successfully used in many applications (Banfield and Raftery, 1983; Fraley and Raftery, 2003). PHDs are an extension of the exponential distributions and exhibit many of their favourable properties such as PHDs have memoryless property, i.e., they are defined only on the nonnegative real numbers (Fackrell, 2009). PHDs can realistically model the process of a patient's journey through different stages of care as a finite state continuous time Markov chain (Fackrell, 2009). While PHDs provide an intuitive description of the patient pathways followed, the GMD is generally better understood by medical practitioners and therefore may be more acceptable. Coxian phase type distributions (C-PHDs) are a special type of PHDs, in which a process starts only in the first transient state and process is eventually absorbed into the absorbing state after sequential transitions through a number of transient states. C-PHDs provide a simple interpretation of fit for the length of stay, and data, require estimating a lot less parameters among all PHDs and also have many other advantages over other types of PHDs (Fackrell, 2009; Marshall and Zenga, 2009). PHDs can also be used to approximate any nonnegative distribution arbitrarily closely (Asmussen, 2003; Fackrell, 2009; Faddy, 1994). Vasilakis and Marshall (2005) fitted the C-PHD to the patients' length of stay data from HES database.

2.1. *Coxian Phase Type Distributions*

Phase type distributions (PHDs) are among popular choices to fit spell length of stay data (Fackrell, 2009). A special type of PHDs are Coxian phase type distributions (C-PHDs) in which a process can start only in the first transient state and only sequential transitions are possible among transient states; transition from any state to the absorbing state is also possible. Fackrell (2009) compares five subclasses of phase type distributions and based on log-likelihoods he identified that the general phase type distributions provide the best fit followed by the Coxian phase type distributions (C-PHDs). However, general phase type distributions are over-parameterized and parameter estimation is difficult (Fackrell, 2009; Marshall and McClean, 2004). On the other hand, Coxian phase type distributions do not present such problems and also provide a simple interpretation of fit for the length of stay data (Fackrell, 2009). We model patient flow in the care system as an n state Markov process (Fig. 1) with Coxian phase type distributions (Cox, 1955; Marshall and McClean, 2004; McClean *et al.*, 2007). A patient can be admitted to the care system only in the first state (state 1). Sequential transitions are possible from any state k (where $k = 1, 2, \ldots, n$) to the next state $k+1$ with a transition rate λ_k. Also transition is possible from any state k to the absorbing state $n+1$ with a transition rate μ_k. The absorbing state represents the event discharge or death of the patient. The time spent in the hospital before discharge or death has the probability density function:

$$f(t) = \mathbf{p}\big(\exp\{\mathbf{Q}t\}\big)\mathbf{q}, \tag{1}$$

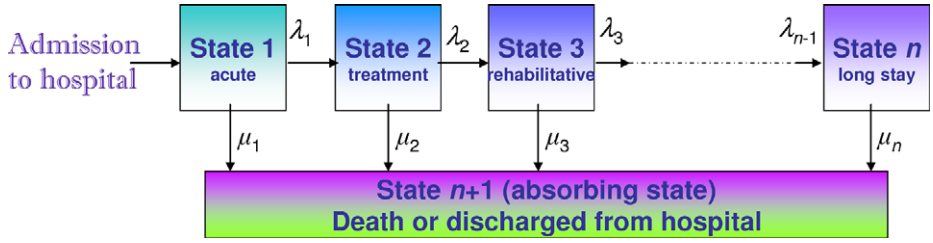

Fig. 1. Stroke care system modelled as an n state Markov process with Coxian phase type distribution.

where the row vector **p**, the initial state probability distribution is defined as:

$$\mathbf{p} = (\,1\;0\;0\;\ldots\;0\;0\,); \tag{2}$$

the transition matrix **Q** is defined as

$$\mathbf{Q} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \ldots & 0 & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & 0 & 0 \\ 0 & 0 & 0 & -(\lambda_{n-1}+\mu_{n-1}) & \lambda_{n-1} & \\ 0 & 0 & 0 & \ldots & 0 & -\mu_n \end{pmatrix}, \tag{3}$$

and the column vector **q** represents absorption probabilities and is defined as

$$\mathbf{q} = (\,\mu_1\;\;\mu_2\;\;\ldots\;\;\mu_{n-2}\;\;\mu_n\,)^{\mathrm{T}}. \tag{4}$$

The likelihood function is defined as follows (Marshall and McClean, 2004):

$$l = \prod_{i=1}^{N} \left(\mathbf{p}\exp\{\mathbf{Q}t_i\}\mathbf{q}\right), \tag{5}$$

where N is the total number of patients in the care system and t_i is the spell length of stay of a patient i $(i = 1, 2, 3, \ldots, N)$. It is more convenient to work with log likelihood function which can be defined as:

$$L = \sum_{i=1}^{N} \left(\log\left(\mathbf{p}\exp\{\mathbf{Q}t_i\}\mathbf{q}\right)\right). \tag{6}$$

This can also be written as:

$$L = \sum_{i=1}^{N} \left(\log(f(t_i))\right). \tag{7}$$

where

$$f(t_i) = \mathbf{p}\big(\exp\{\mathbf{Q}t_i\}\big)\mathbf{q}. \tag{8}$$

This order n Coxian phase type fit of spell length of stay data has $df = 2n - 1$ free parameters (degrees of freedom) to be estimated. We used a freely available downloadable package EMpht (Asmussen *et al.*, 1996) developed by Asmussen *et al.* (1996) and Olsson (1996), which implements maximum likelihood parameter estimation using the expectation-maximization (EM) algorithm. It is a very efficient tool for parameter estimation for phase type distribution when fitted to data (Fackrell, 2009). However, the problem with the tool is that it requires us to specify the number of iterations in advance, which is very difficult to estimate and risking over-fitting. We modified the tool to overcome these limitations. We implemented a stopping criteria based on the improvement in log-likelihood for the fit.

2.2. *Gaussian Mixture Distributions*

Gaussian mixture distributions comprise of a number of normally distributed components. Each component has 3 parameters: the mixing proportion, mean and variance. The probability distribution function for Gaussian mixture distribution with n components is therefore:

$$f(t) = \sum_{k=1}^{n} \alpha_k \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(t-\mu_k)^2}{2\sigma_k^2} \right) \right), \tag{9}$$

and log-likelihood function is:

$$L = \sum_{i=1}^{N} \left(\log \left(\sum_{k=1}^{n} \alpha_k \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(t_i-\mu_k)^2}{2\sigma_k^2} \right) \right) \right) \right). \tag{10}$$

This can also be written as:

$$L = \sum_{i=1}^{N} \big(\log \big(f(t_i) \big) \big), \tag{11}$$

where

$$f(t_i) = \sum_{k=1}^{n} \alpha_k \left(\frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(t_i-\mu_k)^2}{2\sigma_k^2} \right) \right). \tag{12}$$

The degrees of freedom (df) of the distribution with n components are:

$$df = 3n - 1. \tag{13}$$

This model was fitted to the patients' length of stay data from the HES database using the statistical toolbar of MATLAB software which implements maximum likelihood parameter estimation using the expectation-maximization (EM) algorithm (McLachlan and Peel, 2000).

3. Tree Construction

In this section, we will describe the criteria used for phase type survival tree construction.

3.1. *The Splitting Criteria*

Tree construction can be achieved by recursively partitioning into sub groups by one of the covariates based on some splitting criteria maximizing either within node homogeneity or between node separation (Gao *et al.*, 2004). We used splitting criteria to maximize within node homogeneity based on improvement of log-likelihood functions (David and Anderson, 1989). A covariate a can have any of the l values such that

$$N = N_{a1} + N_{a2} + \cdots + N_{al} = \sum_{i=1}^{l} N_{ai}. \tag{14}$$

Therefore (7) (or (11)) can also be written as follows:

$$\begin{aligned} L = \sum_{j=1}^{l} \sum_{i=1}^{N_{aj}} \big(\log\big(f(t_{iaj})\big)\big) &= \sum_{i=1}^{N_{a1}} \big(\log\big(f(t_{ia1})\big)\big) \\ &+ \sum_{i=1}^{N_{a2}} \big(\log\big(f(t_{ia2})\big)\big) + \cdots + \sum_{i=1}^{N_{al}} \big(\log\big(f(t_{ial})\big)\big) \end{aligned} \tag{15}$$

or

$$L = L_{a1} + L_{a2} + \cdots + L_{al} = \sum_{i=1}^{l} L_{ai}. \tag{16}$$

Covariate a splits the dataset into l subgroups and each subgroup is separately fitted to the distribution where the total log-likelihood

$$\begin{aligned} L = \sum_{j=1}^{l} \sum_{i=1}^{N_{aj}} \big(\log\big(f^{(aj)}(t_{iaj})\big)\big) &= \sum_{i=1}^{N_{a1}} \big(\log\big(f^{(a1)}(t_{ia1})\big)\big) \\ &+ \sum_{i=1}^{N_{a2}} \big(\log\big(f^{(a2)}(t_{ia2})\big)\big) + \cdots + \sum_{i=1}^{N_{al}} \big(\log\big(f^{(al)}(t_{ial})\big)\big). \end{aligned} \tag{17}$$

If each subgroup is separately fitted to the Coxian phase type distribution then the likelihood of each fit is

$$f^{(aj)}(t_{iaj})_{\mathrm{PHD}} = \mathbf{p}^{(aj)}\big(\exp\{\mathbf{Q}^{(aj)}t_{iaj}\}\big)\mathbf{q}^{(aj)}. \tag{18}$$

If each subgroup is separately fitted to the Gaussian mixture distribution then the likelihood of each fit is

$$f^{(aj)}(t_{iaj})_{\mathrm{GMM}} = \sum_{k=1}^{n_{aj}} \alpha_k^{(aj)} \left(\frac{1}{\sqrt{2\pi(\sigma_k^{(aj)})^2}} \exp\left(- \frac{(t_{iaj} - \mu_k^{(aj)})^2}{2(\sigma_k^{(aj)})^2} \right) \right). \tag{19}$$

In other words the total log-likelihood is the sum of individual log-likelihoods of each sub-group partitioned by covariate a.

3.2. *The Selection Criteria*

Cross-validation, bootstrap re-sampling and other popular pruning techniques are extremely expensive for large datasets (Gao *et al.*, 2004). Therefore, we have used a simpler approach to determining if a node is a terminal node. If it is not then we select the best possible partition by exploring all possible splits. A terminal node is the node at which within node homogeneity cannot significantly be improved by any possible split. At each node we will apply one covariate at a time and record the total log-likelihood for partitioning by that covariate. Then we will repeat this with other covariates. The covariate which maximizes the total log-likelihood of sub-groups is determined and $L_{\max}$ is calculated as follows.

$$L_{\max} = \max(L_a, L_b, \ldots, L_l). \tag{20}$$

Now we compare this log-likelihood with the log-likelihood of the node before partition and calculate the value of the chi-square statistic $\chi^2_{(df)}$:

$$\chi^2_{(df)} = -2(L_{\max} - L_P), \tag{21}$$

where L_P is the log-likelihood of the node before partition (i.e., the parent node) and degrees of freedom

$$df = df_{\max} - df_P, \tag{22}$$

where $df_{\max}$ is the sum of the degrees of freedom of each of the subgroups partitioned by the log-likelihood maximizing covariate and df_P is the degree of freedom of the node before partition (i.e., the parent node). Here, the null hypothesis is that the node is a terminal node (i.e., no split provides significant improvement in the likelihood ratio function). We used significance level 0.05, i.e., ($\chi^2_{(df)}$ $(p < 0.05)$) to determine if the node is a terminal node. The choice of this significant level is strictly arbitrary and other values (such as 0.1, 0.025, 0.01 or 0.001) can also be used.

4. Running Examples

To illustrate the phase type survival tree and Gaussian distribution survival tree methods for clustering patients according to their hospital length of stay, we used the dataset available from the English Hospital Episode Statistics (HES) database representing the first episode of care of 105,765 patients with a stroke related code anywhere in their diagnosis chain and discharged between April 1st 1994 and March 31st 1995 from all English hospitals (Vasilakis and Marshall, 2005). All patients were aged 65 or over. No information that identified individual patients was supplied. The range of LOS is 0 days to 4906 days, average LOS is 14.2915 days and standard deviation is 52.04 days (Abbi *et al.*, 2008; Vasilakis and Marshall, 2005).

For this application we identified one continuous covariate, i.e., patients' age at the time of admission to hospital and two categorical covariates, i.e., patient gender and type of stroke diagnosed. For the continuous covariate we used cut-points that divide patients into three almost equal subgroups. For the categorical covariate patient gender, HES dataset has four different values 1 for male, 2 for female, 3 and 4 for other or unspecified. Values 3 and 4 do not have prognostic significance. Therefore, we discarded daughter nodes created by patient gender covariates having value 3 or 4. The value of the covariate type of stroke diagnosed is determined by the presence of a particular ICD-9 code (World Health Organisation, 1977) anywhere in the diagnostic chain. It can have any of the 4 values. Hemorrhagic Stroke (ICD-430-ICD-432), Ischemic Stroke (ICD-433, ICD-434, ICD-436, ICD-437), Transient Ischemic Attack (TIA; ICD-435) and other strokes (ICD-438).

4.1. *Phase Type Survival Tree*

Figure 2 is the schematic diagram of the Phase type survival tree we constructed. Table 1 lists the nodes and the possible splits of the tree we constructed. Bold faced covariates represent the splits selected for creating daughter nodes. Node 9, 10, 11 and 12 are nodes created through splitting node 8 by diagnosis covariate.

The total gain in the homogeneity by clustering into leaf nodes in terms of log-likelihood is the difference between root node log-likelihood before clustering and the total log-likelihood of the leaf nodes:

$$G_{Total} = -\big(L_{root} - (L_4 + L_5 + L_6 + L_7 + L_9 + L_{10} + L_{11} + L_{12}) - L_{discard}\big), \tag{23}$$

where $L_{discard}$ is the log-likelihood of the sub-groups which were discarded (with covariate patient's gender value '3' or '4'). The total log-likelihood of leaf nodes of the tree is -368408.967 with the total gain in log-likelihood is 3793.631635 with 35 extra free parameters ($p = 1$).

Table 1

Phase type survival tree construction for HES database on stroke-related patients

Node	Covariate	Covariate value	Number of patients	Loglikelihood	Number of phases	Value of Chi square statistic $\chi^2_{(df)}$	Additional degree of freedom (df)	Chi-square probability (p) of selected split
All	Complete dataset		105765	−372202.5986	14			
1	Gender	Male	47136	−159483.7033	7	942.438984	20	< 0.000001
(Root node)		Female	58109	−210420.5992	9			
		Unspecified	520	−1827.076608	9			
	Age	**Age <= 73**	**34995**	**−108847.6535**	**7**	**7297.3144**	**24**	**< 0.000001**
		73 < Age < 82	**35393**	**−124625.1882**	**11**			
		Age >= 82	**35377**	**−135081.0997**	**9**			
	Diagnosis	Hemorrhagic	5593	−19510.25135	9	202.79572	39	< 0.000001
		Ischemic	67190	−236434.3777	8			
		TIA	11196	−39030.51958	9			
		Other	21786	−77126.05211	9			
2	**Gender**	**Male**	**11720**	**−43505.86672**	**4**	**120.255304**	**4**	**< 0.000001**
(Age >= 82)		**Female**	**23485**	**−90841.97838**	**4**			
		Unspecified	**172**	**−673.126948**	**4**			
	Diagnosis	Hemorrhagic	1860	−7111.186087	4	−79.630374	11	–
		Ischemic	22605	−86486.26874	4			
		TIA	3777	−14364.33872	4			
		Other	7135	−27159.12134	4			
3	**Gender**	**Male**	**16419**	**−56342.3129**	**4**	**136.968716**	**0**	**< 0.000001**
(73 < Age <		**Female**	**18808**	**−67631.89135**	**4**			
82)		**Unspecified**	**166**	**−582.499592**	**4**			
	Diagnosis	Hemorrhagic	1871	−6611.551198	4	−69.826996	9	–
		Ischemic	22351	−78619.50522	4			
		TIA	3697	−12928.67402	5			
		Other	7474	−26500.37126	4			
4	Gender	Male	18997	−58518.12493	4	−365.790274	8	–
(Age <= 73)		Female	15816	−49962.72347	4			
		Unspecified	182	−549.700237	4			
	Diagnosis	Hemorrhagic	1862	−5634.446937	5	−292.196714	21	–
		Ischemic	22234	−69026.23065	5			
		TIA	3722	−11338.31477	4			
		Other	7177	−22994.7595	5			
5	Diagnosis	Hemorrhagic	593	−2180.100904	4	−33.38052	19	–
(Age >= 82)		Ischemic	7387	−27518.89464	4			
Male		TIA	1246	−4636.007606	4			
		Other	2494	−9187.55383	3			
6	Diagnosis	Hemorrhagic	1259	−4888.606315	4	−118.438468	21	–
(Age >=82)		Ischemic	15100	−58543.49525	4			
Female		TIA	2511	−9641.868259	4			
		Other	4615	−17827.22779	4			
7	Diagnosis	Hemorrhagic	860	−2933.985556	4	7.774826	21	0.996027
(73 < Age <		Ischemic	10191	−34900.37992	4			
82 Male)		TIA	1740	−5945.469171	4			
		Other	3628	−12558.59084	4			
8	**Diagnosis**	**Hemorrhagic**	**998**	**−3622.623753**	**4**	**32.72485**	**21**	**0.049369**
(73 < Age <		**Ischemic**	**12061**	**−43332.41141**	**4**			
82 Female)		**TIA**	**1934**	**−6883.738102**	**4**			
		Other	**3815**	**−13776.75566**	**4**			

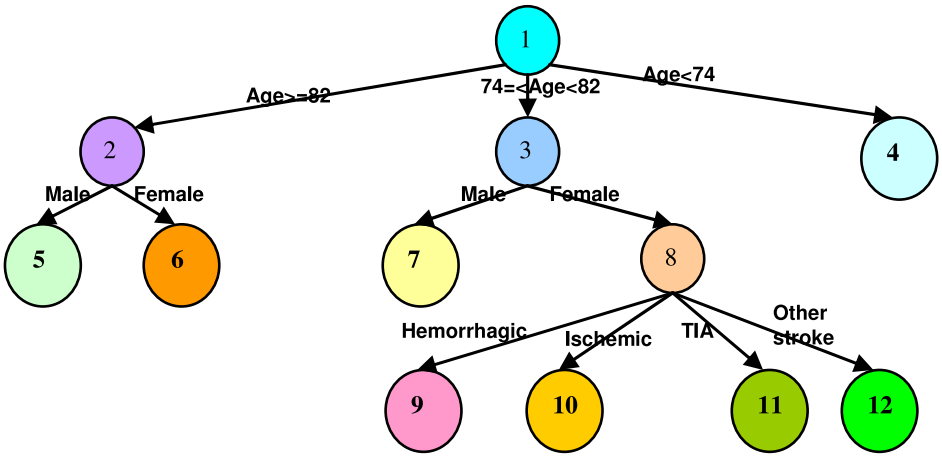

Fig. 2. Phase-type survival tree for HES database on stroke-related patients.

Table 2

Gaussian mixture survival tree construction for HES database on stroke-related patients

Node	Covariate	Covariate value	Number of patients	Log-likelihood likelihood	Number of phases	Value of Chi square statistic $\chi^2_{(df)}$	Additional degree of freedom (df)	Chi-square probability (p) of selected split
All	1		**105765**	**−368687.3**	**7**	–	–	–
1 (Root node)	Gender	Male	47136	−162739.6	6	−18932.654	28	–
		Female	58109	−213549.5	6			
		Unspecified	520	−1864.527	5			
	Age	Age <= 73	34995	−120869.3	3	−30824.8	25	–
		73 < Age < 82	35393	−126709	6			
		Age >= 82	35377	−136521.4	7			
	Diagnosis	Hemorrhagic	5593	−17863.81	5	−19780.18	36	–
		Ischemic	67190	−242176.9	5			
		TIA	11196	−40059.46	4			
		Other	21786	−78477.22	6			

4.2. *Gaussian Mixture Survival Tree*

Similar to the phase type survival tree we can construct a Gaussian distribution survival tree for the HES dataset. However, while constructing the tree, no covariate improved the within node homogeneity at the root node.

5. Mixture Distribution Survival Trees

We can achieve more improvement in the likelihood function by introducing more flexibility in the tree construction where different nodes in the tree can be approximated by distinct types of mixture distributions. This results in the improved within node homo-

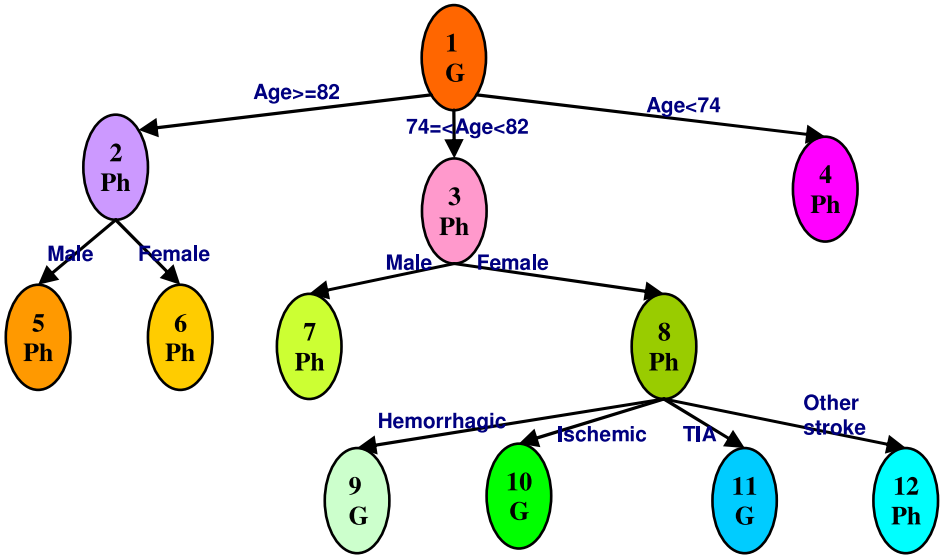

Fig. 3. Mixed distribution survival tree for the HES database.

geneity. The tree is constructed by recursively partitioning the dataset into subgroups using one of the covariates. Each subgroup is separately fitted to both C-PHD and GMD. We use the same splitting and selection criteria as discussed in Section 3. The distribution fitting with maximum log-likelihood is selected for that subgroup. A split which maximizes node homogeneity by providing maximum positive improvement in the total log-likelihood (i.e., the sum of individual log-likelihoods of each subgroup split by the covariate) is selected to grow the tree. If at a node, there is no split providing improvement in the log-likelihood, the node is designated as a terminal node.

Figure 3 is the schematic representation of the final mixed distribution survival tree for the length of stay data from the HES database. A node with 'Ph' is modelled by PHD while a node with 'G' is modelled by GMD, i.e., root node (1), node 9, node 10 and node 11 are modelled by GMDs and node 2, node 3, node 4, node 5, node 6, node 7, node 8 and node 12 are modelled by PHDs. Table 5 presents the tree construction process listing nodes of the tree and possible splits of these nodes. Bold faced covariates were selected for splitting the parent node. At each node, the distribution having log-likelihoods written in italic is selected.

The total gain (G_{Total}) in the homogeneity by clustering into leaf nodes in terms of log-likelihood can be calculated using equation 23. The log-likelihood of the root node (modelled by GMD) is -372202.6. There is significant improvement in the log-likelihood after clustering with mixture distribution survival tree and the total log-likelihood of leaf nodes is -367119.8101 which is improved by 1289.1569 in comparison with phase type survival tree. The total gain in the within node homogeneity by clustering into leaf nodes using in mixture distribution survival tree is 1567.488535. However, if the root node were approximated by phase type distribution, then the total gain in the within

Table 3

Mixed distribution survival tree construction for the LoS data from HES database

Parent nodes	Covariate	Covariate value	PHD loglikelihood	PHD Number of phases	GMD loglikelihood	GMD Number of phases	Distribution selected	Value of Chi square statistic $\chi^2_{(df)}$	Additional degree of freedom (df)	Chi-square probability (p) of selected split
All	All	–	−372202.6	14	**−368687.3**	7	GMD	–	–	–
1 (Root node)	Sex	Male	−159483.7	7	−162739.6	6	PHD	−6088.15	27	–
		Female	−210420.6	9	−213549.5	6	PHD			
		Unspecified	−1827.0766	9	−1864.527	5	PHD			
	Age	**Age <= 73**	**−108847.65**	**7**	−120869.3	3	PHD	266.72	31	**< 0.000001**
		73 < Age < 82	**−124625.19**	**11**	−126709	6	PHD			
		Age >= 82	**−135081.1**	**9**	−136521.4	7	PHD			
	Diagnosis	hemorrhagic	−19510.251	9	−17863.81	5	PHD	−6827.81	46	–
		ischemic	−236434.38	8	−242176.9	5	PHD			
		TIA	−39030.52	9	−40059.46	4	PHD			
		Other	−77126.052	9	−78477.22	6	PHD			
2 (Age >= 82)	**Sex**	**Male**	**−43505.867**	**4**	−43985.64	8	PHD	120.2561	4	**< 0.000001**
		Female	**−90841.978**	**4**	−91726.43	7	PHD			
		unspecified	**−673.12695**	**4**	−680.2635	4	PHD			
	Diagnosis	hemorrhagic	−7111.1861	4	−7200.211	6	PHD	−79.6302	11	–
		ischemic	−86486.269	4	−87407.6	7	PHD			
		TIA	−14364.339	4	−14497.73	7	PHD			
		Other	−27159.121	4	−27403.26	8	PHD			
3 (73 < Age < 82)	**Sex**	**Male**	**−56342.313**	**4**	−57451.84	6	PHD	136.9728	0	**< 0.000001**
		Female	**−67631.891**	**4**	−69098.01	5	PHD			
		unspecified	**−582.49959**	**4**	−587.5177	5	PHD			
	Diagnosis	hemorrhagic	−6611.5512	4	−6694.673	6	PHD	−69.8224	9	–
		ischemic	−78619.505	4	−80453.24	5	PHD			
		TIA	−12928.674	5	−13604.29	4	PHD			
		Other	−26500.371	4	−26902.44	6	PHD			
4 (Age <= 73)	Sex	Male	−58518.125	4	−64672.55	3	PHD	−365.796	8	–
		Female	−49962.723	4	−55547.53	3	PHD			
		unspecified	−549.70024	4	−598.0513	3	PHD			
	Diagnosis	hemorrhagic	−5634.4469	5	−6403.329	3	PHD	−292.206	21	–
		ischemic	−69026.231	5	−76699.52	3	PHD			
		TIA	−11338.315	4	−12587.48	3	PHD			
		Other	−22994.76	5	−24155.37	4	PHD			
5 (Age >= 82 Male)	Diagnosis	hemorrhagic	−2180.1009	4	−2211.344	6	PHD	−33.3806	19	–
		ischemic	−27518.895	4	−27781.94	7	PHD			
		TIA	−4636.0076	4	−4684.628	6	PHD			
		Other	−9187.5538	3	−9263.08	6	PHD			
6 (Age >= 82 Female)	Diagnosis	hemorrhagic	−4888.6063	4	−4921.049	7	PHD	−118.439	21	–
		ischemic	−58543.495	4	−59058.94	7	PHD			
		TIA	−9641.8683	4	−9711.659	7	PHD			
		Other	−17827.228	4	−17991.41	6	PHD			
7 (73 <Age < 82 Male)	Diagnosis	hemorrhagic	−2933.9856	4	−2961.871	7	PHD	7.7744	21	0.996029
		ischemic	−34900.38	4	−35563.87	5	PHD			
		TIA	−5945.4692	4	−6098.857	5	PHD			
		Other	−12558.591	4	−12792.26	7	PHD			
8 (73 <Age < 82 Female)	**Diagnosis**	**hemorrhagic**	−3622.6238	**4**	**−3588.648**	**8**	GMD	2611.038	73	**< 0.000001**
		ischemic	−43332.411	**4**	**−42399.08**	**8**	GMD			
		TIA	−6883.7381	**4**	**−6561.888**	**6**	GMD			
		Other	**−13776.756**	**4**	−13956.05	6	PHD			

node homogeneity by clustering into leaf nodes using in mixture distribution survival tree would have been 5082.788535.

6. Discussion

This paper has presented two novel types of survival trees, phase type survival tree, and mixture distribution survival tree, which can effectively be used for prognostication of survival data and for clustering survival data into groups of patients following homogeneous patient pathways. (A patient pathway can be defined as a sequence of partially observable states followed by a patient before absorption, i.e., before discharge or death). These are effective tools for determining the relationship between input covariates and outcome measures and their interrelations. They provide understanding of heterogeneity of patient pathways stratified by covariates representing patient characteristics such as age, gender and diagnosis. We can also use these models to estimate the length of stay of a patient based on his/her characteristics available at the time of admission. These survival tree based models together with the model proposed by Garg *et al.* (2010) can be used for better estimations of resource requirements and cost of care as it considers the effects of individual cluster of patients, their interactions in the whole care unit and thus the effect of demographic changes in the patient population (Garg, 2009).

The results in the above section demonstrate that using mixed distribution survival trees we can have more homogeneous clusters and improvement in the total gain (G_{Total}) in the homogeneity. We used log likelihood ratio based slitting criteria for node split and selection. We can also use other splitting criteria such as Akaike Information Criterion (AIC; Akaike, 1974) or Bayesian Information Criterion (BIC; Schwartz, 1978).

Mixed distribution survival trees can also be constructed using other distributions such as General phase type, Weibull, Gamma, Beta, Bernoulli, binomial, multinomial, Laplacian, exponential, Wishart, Dirichlet, Rayleigh, Pareto distributions etc (Nielsen and Garcia, 2009). However, in this paper we used only Coxian PHDs and GMDs as they represent most of the distributions from the class of exponential distributions. Exponential and hyper exponential distributions are types of phase type distributions (Fackrell, 2009). For shape parameter value $= 1$, both Gamma distributions and Weibull distributions with location parameter value $= 0$, reduces to the exponential distributions (Martinez and Martinez, 2007). Also PHDs can be used to approximate any nonnegative distribution arbitrarily closely (Asmussen, 2003; Fackrell, 2009).

Theoretically length of stay can be defined as a continuous variable while in practice it is recorded as a discrete number (integer or with a fixed number of decimal places). However, for a large sample population with few ties, hospital length of stay can still be modelled as a continuous time random variable. In the healthcare literature, many researchers have modelled hospital length of stay as a continuous random variable (Abbi *et al.*, 2008; El-darzi *et al.*, 2009; Fackrell *et al.*, 2009; Marshall *et al.*, 2007; Marshall and McClean, 2004; Marshall and Zenga, 2009; McClean *et al.*, 2007; Riihimäki *et al.*, 2010).

7. Conclusion

In this paper we illustrate how phase type survival trees and mixture distribution survival tree can be used to cluster, identify and quantify the significance and effects of various covariates (patient characteristics such as age, gender, disease etc.) and their interaction in prediction of patient's length of stay in hospital. We have used splitting criteria based on improvement of log-likelihood functions. The mixture distribution survival tree provides better improvement in the likelihood function and in the within node homogeneity. As future work we will determine the effect of using other splitting criteria to develop more efficient clustering. Also for continuous covariates we will develop an automated algorithm which can be used to decide optimum cut points. Presently we are working on using our survival tree based models together with the model proposed by Garg *et al.* (2010) for better estimation of resource requirements and cost of care.

Acknowledgement. The authors acknowledge support for this work from the Engineering and Physical Sciences Research Council (Grant References EP/E019900/1 and GR/S29874/01). Any views or opinions presented herein are those of the authors and do not necessarily represent those of RIGHT or MATCH, their associates or their sponsors.

References

Abbi, R., El-Darzi, E., Vasilakis, C., Millard, P. (2008). A Gaussian mixture model approach to grouping patients according to their hospital length of stay. In: *Proceedings – IEEE Symposium on Computer-Based Medical Systems*, pp. 524–529.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edn. Springer, Berlin.

Asmussen, S., Nerman, O., Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23, 419–441.

Banfield, J.D., Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Wadworth and Brooks/Cole, Monterey.

Clarke, J., West, M. (2007). Bayesian Weibull tree models for survival analysis of clinico-genomic data. *Statistical Methodology*, 5(3), 238–262.

Cox, D.R. (1955). A use of complex probabilities in the theory of stochastic processes. *Proceedings of the Cambridge Philosophical Society*, 51, 313–319.

Davis, R., Anderson, J. (1989). Exponential survival trees. *Statistics in Medicine*, 8, 947–962.

El-darzi, E., Abbi, R., Vasilakis, C., Gorunescu, F., Gorunescu, M., Millard, P. (2009). Length of stay-based clustering methods for patient grouping. *Studies in Computational Intelligence*, 189, 39–56.

Fackrell, M. (2009). Modelling healthcare systems with phase-type distributions. *Health Care Management Science*, 12, 11–26.

Faddy, M.J. (1994). Examples of fitting structured phase-type distributions. *Applied Stochastic Models and Data Analysis*, 10, 247–55.

Fan, J., Nunn, M.E., Su, X. (2009). Multivariate exponential survival trees and their application to tooth prognosis. *Computational Statistics and Data Analysis*, 53(4), 1110–1121.

Fraley, C., Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611–631.

Gao, F., Manatunga, A.K., Chen, S. (2004). Identification of prognostic factors with multivariate survival data. *Computational Statistics & Data Analysis*, 45, 813–824.

Garg, L. (2009). Unified modelling for care of the elderly. PhD thesis, School of Computing and Information Engineering, Faculty of Computing and Engineering, University of Ulster.

Garg, L., McClean, S.I., Meenan, B.J., Millard, P.H. (2010). A non-homogeneous discrete time Markov model for admission scheduling and resource planning in a care system. *Healthcare Management Science*, 13(2), 155–169; doi:10.1007/s10729-009-9120-0.

Marshall, A.H., McClean S.I. (2004). Using coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. *Healthcare Management Science*, 7, 285–289.

Marshall, A.H., Zenga, M. (2009). Recent developments in fitting coxian phase-type distributions in healthcare. In: Sakalauskas, L., Skiadas, C., Zavadskas, E.K. (Eds.), *The XIIIth International Conference "Applied Stochastic Models and Data Analysis" (ASMDA-2009)*, Vilnius, Lithuania, Selected papers. ISBN 978-9955-28-463-5, pp. 482–485.

Marshall, A.H., Burns, M.L., Shaw, B. (2007). Patient activity in hospital using discrete conditional phase-type (DC-PH) models. In: Skiadas, C.H. (Ed.), *Recent Advances in Stochastic Modelling & Data Analysis*. World Scientific, Singapore, pp. 154–161.

Martinez, W.L., Martinez, A.R. (2007). *Computational Statistics Handbook with Matlab*, 2nd edn., Chapman & Hall/CRC, London/Boca Raton.

McClean, S.I., Garg, L., Meehan, B., Millard, P.H. (2007). Non-homogeneous Markov models for performance monitoring in healthcare. In: Skiadas, C.H. (Ed.), *Recent Advances in Stochastic Modelling and Data Analysis*, pp. 146–153.

McLachlan, G., Peel, D. (2000). *Finite Mixture Models*. Wiley, Hoboken.

Nielsen, F., Garcia, V. (2009). *Statistical Exponential Families: A digest with Flash Cards*, v1.0, informal publication. CoRR abs/0911.4863 avialable on internet at `http://arxiv.org/PS_cache/arxiv/pdf/0911/0911.4863v1.pdf` (Accessed on April 22nd, 2010).

Olsson, M. (1996). Estimation of phase-type distributions from censored data. *Scandinavian Journal of Statistics*, 23, 443–460.

Riihimäki, J., Sund, R., Vehtari, A. (2010). Analysing the length of care episode after hip fracture: a nonparametric and a parametric Bayesian approach. *Health Care Management Science*, 13(2), 170–181.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.

Šutienė, K., Makackas, D., Pranevičius, H. (2010). Multistage K-means clustering for scenario tree construction. *Informatica*, 21(1), 123–138.

Vasilakis, C., Marshall, A.H. (2005). Modelling nationwide hospital length of stay: opening the black box. *Journal of the Operational Research Society*, 56, 862–869

World Health Organisation (1977). *International Classification of Diseases*, 9 revision (ICD-9). WHO, Geneva.

Žilinskas, A., Žilinskas, J. (2010). Interval arithmetic based optimization in nonlinear regression. *Informatica*, 21(1), 149–158.

L. Garg received his first degree in electronics and communication engineering from Barkatullah University, Bhopal, India and then obtained PGD in information technology from ABV-Indian Institute of Information Technology and Management, Gwalior, India followed by a PhD on healthcare modelling at the University of Ulster, UK. He is currently research associate in computational mathematics at the University of Ulster, UK. His research interests are mathematical modelling, stochastic modelling, artificial intelligence, operation research, management science and data mining.

S. McClean is professor of mathematics at the University of Ulster. Her main research interests are in statistical modelling and optimisation, particularly for healthcare planning, and computer science, particularly databases, sensor technology and telecommunications. She is currently grantholder on over Ł7M worth of funding, mainly from the EPSRC and other government sources. Sally is a fellow of the Royal Statistical Society, and a past president of the Irish Statistical Association. She is a recipient of Ulster's Senior Distinguished Research Fellowship.

B.J. Meenan is professor biomedical materials at the University of Ulster and is head of the Biomaterials & Tissue Engineering Research Group (BTERG) within the Nanotechnology & Integrated Bioengineering Centre (NIBEC). He is director of University of Ulster Hub of the EPSRC funded MATCH (Multidisciplinary Assessment of Technology Centre for Healthcare) Innovation Manufacturing Research Centre (IMRC). Professor Meenan's research interests are directed toward the development of functional biomaterials for medical applications with particular emphasis on the promotion of interactions on the submicron to nano-metre scale for applications in bio-responsive medical implants, tissue engineering and regenerative medicine. He also active in the area of evaluation processes for innovative healthcare technologies and new assessment methods to predict their clinical utility.

P. Millard is the emeritus professor of geriatrics at St. George's Hospital Medical School. Past president of the British Geriatrics Society. President UK Nosokinetics Group. Editor Nosokinetics News Research interests relate to the ethical, organisational and educational principles that should underpin the provision of health and social care for an ageing population. His MD (1989) thesis developed a theoretical model of flow of patients through geriatric medical beds. His PhD (1993) tested the practical application of Harrison's 1991 mathematical solution. That research has now led to a small world-wide effort involving mathematicians, decision scientists and health care planners developing new ways of measuring, forecasting and costing the dynamic interactions between acute, rehabilitative and long-stay care in hospital and in the community.

Fazinio tipo ir skirstinių mišinių išgyvenamumo medžiai pacientų ligoninėse praleistų dienų skaičiui klasteriuoti

Lalit GARG, Sally McCLEAN, Brian MEENAN, Peter MILLARD

Klinikų tyrėjai, sveikatos profesionalai ir vadybininkai dažnai ieško kriterijų pacientams sugrupuoti į kliniškai reikšmingas grupes, siekiant įvertinti tikėtiną ligoninėje praleidžiamų dienų skaičių. Šiame straipsnyje yra pasiūlyti du išgyvenamumo medžių tipai, būtent, fazinio tipo išgyvenamumo medžiai ir skirstinių mišinių išgyvenamumo medžiai, išplėtojant ankstesniuosius darbus, skirtus eksponentiniams išgyvenamumo merdžiams. Medžiai pritaikomi pacientams suklasteriuoti pagal praleistų logoninėje dienų skaičių, atsižvelgiant į amžių, lytį ir pirminės diagnozės kodą, pasinaudojant tikėtinumo santykiu optimaliam sugrupavimui nustatyti. Metodui iliustruoti yra pasinaudojama duomenimis iš Anglijos ligoninių epizodų statistikos (HES) duomenų bazės apie pacientus su auglių susirgimų diagnoze, vyresnius negu 65 ir išrašytus iš Anglijos ligoninių per vienerius metus.