# The extended mixture distribution survival tree based analysis for clustering and patient pathway prognostication in a stroke care unit

Lalit Garg, Sally McClean, Maria Barton
School of Computing and Information Engineering,
University of Ulster, Coleraine, Northern Ireland, UK

Brian Meenan
School of Engineering,
University of Ulster, Jordanstown, Northern Ireland, UK

Ken Fullerton
School of Medicine, Dentistry and Biomedical Sciences,
Queen's University, Belfast, Northern Ireland, UK

***Abstract*- In our previous work we proposed a special class of survival distribution called Mixture distribution survival trees, which are constructed by approximating different nodes in the tree by distinct types of mixture distributions to achieve more improvement in the likelihood function and thus the improved within node homogeneity. We proposed its applications in modelling hospital length of stay and clustering patients into clinically meaningful patient groups, where partitioning was based on covariates representing patient characteristics such as gender, age at the time of admission, and primary diagnosis code. This paper proposes extended Mixture distribution survival trees and demonstrates its applications to patient pathway prognostication and to examine the relationship between hospital length of stay and/or treatment outcome. 5 year retrospective data of patients admitted to Belfast City Hospital with a diagnosis of stroke is used to illustrate the approach.**

***Keywords*- Stochastic modeling, Survival tree, Length of stay modelling, Prognostication, Clustering, Gaussian mixture distributions, Phase type distributions**

## I. INTRODUCTION

Mixture distribution survival trees [1] are special type of survival trees, which are constructed by approximating different nodes in the tree by distinct types of mixture distributions to achieve more improvement in the likelihood function and thus the improved within node homogeneity. Survival trees can be used as a powerful method for partitioning survival data into clinically meaningful patient groups for prognostication, i.e., for determining importance, effects of various input covariates (such as a patient's characteristics) and their influence on output measures such as patients' survival, their expected length of stay, discharge destination, or treatment outcome [2] [3]. In our previous work [1], we proposed a mixture distribution survival tree based method where tree nodes were approximated using Gaussian mixture distributions and phase type distributions, for into homogeneous groups with respect to their length of stay (LOS) where partitioning was based on covariates representing patient characteristics such as gender, age at the time of admission and primary diagnosis code. This paper extends this approach to patient pathway prognostication i.e. for determining importance and effects of various input covariates such as gender, age at the time of admission and primary diagnosis code on patients' hospital length of stay and to examine the relationship between length of stay in hospital and treatment outcome.

An application of the approach to patient pathway prognostication is illustrated using 5 years' retrospective data [4] for 1985 patients admitted between January 2003 and December 2007 to the Belfast City Hospital with a diagnosis of stroke (hemorrhagic stroke, cerebral infarction, transient ischaemic attack TIA, and stroke unspecified). All patients were discharged between January 9th, 2003 and March 11th 2008. No information that identified individual patients was supplied. Patients were aged between 24 years and 101 years. Patient's lengths of stay range from is 0 days (admitted and discharge on the same day) to 1425 days, mean LOS 29.01 days with 52.84 days standard deviation [4].

## II. MIXTURE DISTRIBUTION SURVIVAL TREE CONSTRUCTION

A survival tree can be constructed by recursively splitting nodes into daughter nodes by one of the covariates based on some splitting criteria either maximizing either within node homogeneity or between node separation [3]. Each daughter node is approximated by both GMD and C-PhD with different set of components. We used splitting criteria to maximize within node homogeneity expressed in terms of Akaike Information criterion (AIC) [5].

AIC = -2*Log likelihood+2*$df$.

Where $df$ is the number of free parameters to be estimated. For nodes modeled by n component (phase) Coxian phase type distribution (C-PhD), $df = 2*n-1$ and for a node modeled by $m$ component Gaussian mixture distribution (GMD), $df = 2*m-1$.

A split with minimum value of AIC is selected. If at a node, there is no split providing positive improvement in the AIC, the node is designated as a terminal node.

We used three covariates gender, age at the time of admission and type of stroke diagnosed. The covariate 'age' has value 'old' for those aged 70 or over and it has value 'young' for those aged below 70 years. Based on the primary ICD-10 diagnosis code [6], patients can have any of the four values (hemorrhagic stroke, cerebral infarction, transient ischaemic attack TIA, and other strokes) for the covariate 'stroke diagnosed'.

Figure 1 is the schematic representation of the final mixture distribution survival tree for the length of stay data on stroke patients from the Belfast City Hospital. The resulting tree has 12 terminal nodes. A node with 'P' is modeled by C-PhD while a node with 'G' is modeled by GMD, i.e., node 9, node 17 and node 19 are modeled by GMDs and all other nodes root node (node 1), node 2, node 3, node 4, node 5, node 6, node 7, node 8, node 10, node 11, node 12, node 13, node 14, node 15, node 16, node 18, node 20 and node 21 are modeled by C-PhDs. Nodes of the tree and possible splits of these nodes are listed in Table 1. Bold faced covariates were selected for splitting the parent node. Nodes which are better fitted by GMD are having AIC shaded yellow. The AIC of root node was 16825.6 and new total AIC of all the terminal nodes of the survival tree is 16397.92 with 427.68 total improvement in AIC.

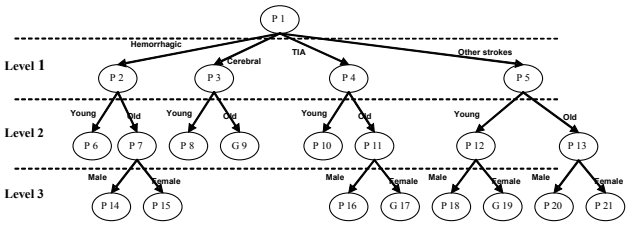


Fig. 1. Mixture distribution survival tree for the length of stay data on stroke patients from the Belfast City Hospital

TABLE I
MIXTURE DISTRIBUTION SURVIVAL TREE CONSTRUCTION (NODES AND POSSIBLE SPLITS)

| Node | Covariate | Covariate value | Number of patients | Mean LoS | Standard deviation (LoS) | Coxian-phase type distribution | | | Gaussian mixture distribution | | | Improvement in AIC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Loglikelihood ($L_{max}$) | Number of phases | AIC | Loglikelihood | Number of phases | AIC | |
| **All** | **Complete dataset** | **Root node** | **1985** | **29.0106** | **52.8382** | **-8407.8007** | **3** | **16825.60** | **-8481.625** | 10 | **17021.25** | - |
| **Level 1** | | | | | | | | | | | | |
| 1 (Root node) | Gender | Male | 933 | 26.5938 | 44.0575 | -3859.8125 | 2 | 7725.625 | -3897.236 | 7 | 7834.47 | 10.2825 |
| | | Female | 1052 | 31.154 | 59.4698 | -4539.8469 | 3 | 9089.694 | -4569.59 | 9 | 9191.18 | |
| | Age | Young | 624 | 19.2564 | 39.1523 | -2316.974 | 2 | 4639.948 | -2342.454 | 7 | 4724.9 | 127.159 |
| | | Old | 1361 | 33.4827 | 57.4932 | -6024.2472 | 3 | 12058.49 | -6047.846 | 9 | 12147.7 | |
| | **Diagnosis** | **Hemorrhagic** | **154** | **33.6039** | **56.4456** | **-659.05019** | **3** | **1328.1** | **-665.5723** | **6** | **1365.14** | **306.1832** |
| | | **Cerebral** | **655** | **36.6611** | **47.6753** | **-2973.8941** | **4** | **5961.788** | **-2980.867** | **6** | **5995.73** | |
| | | **TIA** | **425** | **9.31294** | **19.9516** | **-1298.6262** | **2** | **2603.252** | **-1316.503** | **6** | **2667** | |
| | | **Other** | **751** | **32.5433** | **65.0453** | **-3310.1386** | **2** | **6626.277** | **-3297.443** | **8** | **6640.89** | |
| **Level 2** | | | | | | | | | | | | |
| 2 Hemorrhagic | Gender | Male | 80 | 28.2 | 52.09832 | -317.63202 | 4 | 649.2640 | -328.0819 | 4 | 678.1638 | 12.50299 |
| | | Female | 74 | 39.4459 | 60.254 | -328.16667 | 3 | 666.3333 | -329.4363 | 6 | 692.873 | |
| | **Age** | **Young** | **50** | **24.56** | **55.117** | **-173.39875** | **4** | **360.7975** | **-187.3508** | **3** | **390.702** | **15.65113** |
| | | **Old** | **104** | **37.9519** | **56.561** | **-468.82587** | **4** | **951.6517** | **-477.7462** | **4** | **977.492** | |
| 3 Cerebral | Gender | Male | 302 | 33.70860 | 49.8833 | -1334.898 | 4 | 2683.796 | -1339.464 | 5 | 2706.93 | -2.395836 |
| | | Female | 353 | 39.18697 | 45.5501 | -1635.194 | 3 | 3280.388 | -1639.346 | 5 | 3306.69 | |
| | **Age** | **Young** | **194** | **24.0670** | **42.4506** | **-785.3629** | **3** | **1580.726** | **-793.676** | **5** | **1615.35** | **35.8404** |
| | | **Old** | **461** | **41.961** | **48.787** | **-2173.9068** | **2** | **4353.814** | **-2158.611** | **5** | **4345.22** | |
| 4 TIA | Gender | Male | 207 | 8.7005 | 22.6817 | -607.9547 | 2 | 1221.909 | -637.5377 | 4 | 1297.08 | -1.20391 |
| | | Female | 218 | 9.8945 | 16.9366 | -686.27346 | 3 | 1382.547 | -690.5091 | 5 | 1409.02 | |
| | **Age** | **Young** | **176** | **5.83523** | **11.1641** | **-455.8639** | **2** | **917.7278** | **-465.3673** | **4** | **952.735** | **24.78144** |
| | | **Old** | **249** | **11.7711** | **24.0154** | **-827.3716** | **2** | **1660.743** | **-823.7194** | **8** | **1693.44** | |
| 5 Other strokes | Gender | Male | 344 | 30.7413 | 43.4091 | -1490.577 | 4 | 2995.154 | -1490.229 | 6 | 3014.46 | 8.886526 |
| | | Female | 407 | 34.0663 | 78.7981 | -1808.118 | 2 | 3622.237 | -1864.211 | 3 | 3744.42 | |
| | **Age** | **Young** | **204** | **24.9608** | **43.76126** | **-818.1347** | **4** | **1650.269** | **-819.1983** | **6** | **1672.4** | **36.23604** |
| | | **Old** | **547** | **35.3711** | **71.1697** | **-2466.8858** | **2** | **4939.772** | **-2543.835** | **3** | **5103.67** | |
| **Level 3** | | | | | | | | | | | | |
| 6 Hemorrhagic Young | Gender | Male | 29 | 30.5172 | 69.1114 | -108.83289 | 2 | 223.6658 | -105.682 | 4 | 233.364 | -9.21382 |
| | | Female | 21 | 16.3333 | 22.8126 | -70.172765 | 2 | 146.3455 | -74.17359 | 3 | 164.347 | |
| **7 Hemorrhagic Old** | **Gender** | **Male** | **51** | **26.8823** | **39.2027** | **-211.39224** | **4** | **436.7845** | **-211.0736** | **4** | **444.147** | **2.30463** |
| | | **Female** | **53** | **48.6038** | **67.5821** | **-253.2813** | **2** | **512.5626** | **-245.9886** | **7** | **531.977** | |
| 8 Cerebral Young | Gender | Male | 104 | 24.6731 | 49.2715 | -420.88798 | 2 | 847.776 | -427.2601 | 4 | 876.52 | -3.15333 |
| | | Female | 90 | 23.36667 | 32.9415 | -361.0516 | 4 | 736.1032 | -364.2716 | 4 | 750.543 | |
| 9 Cerebral Old | Gender | Male | 198 | 38.4545 | 49.6696 | -903.58419 | 4 | 1821.168 | -903.381 | 4 | 1828.76 | -0.62756 |
| | | Female | 263 | 44.6008 | 47.9429 | -1259.34 | 2 | 2524.681 | -1251.528 | 6 | 2537.06 | |
| 10 TIA Young | Gender | Male | 88 | 5.7386 | 11.3263 | -224.74588 | 2 | 455.4918 | -232.5108 | 3 | 481.022 | -5.52344 |
| | | Female | 88 | 5.9318 | 10.9988 | -230.8797 | 2 | 467.7595 | -231.1457 | 4 | 484.291 | |
| **11 TIA Old** | **Gender** | **Male** | **119** | **10.8908** | **28.0847** | **-377.7327** | **2** | **761.4654** | **-401.7215** | **3** | **819.443** | **2.615394** |
| | | **Female** | **130** | **12.5769** | **19.5270** | **-444.56883** | **4** | **903.1377** | **-437.3312** | **4** | **896.662** | |
| **12 Other strokes Young** | **Gender** | **Male** | **119** | **30.1092** | **52.7719** | **-493.3325** | **3** | **996.665** | **-486.2055** | **6** | **1006.41** | **0.265422** |
| | | **Female** | **85** | **17.7529** | **24.6624** | **-322.7928** | **3** | **655.5857** | **-318.6695** | **3** | **653.339** | |
| **13 Other strokes Old** | **Gender** | **Male** | **225** | **31.0756** | **37.52** | **-987.5257** | **4** | **1989.051** | **-991.5326** | **4** | **2005.07** | **3.802914** |
| | | **Female** | **322** | **38.3727** | **87.1713** | **-1470.4587** | **2** | **2946.917** | **-1517.217** | **3** | **3050.43** | |

At level 2, for all nodes, the covariate 'age' provided the most significant split while the covariate 'gender' did not provide significant splits for the group of patients with diagnosis cerebral infarction and for the group of patients with diagnosis TIA. For example, among patients with TIA, young patients were most likely to have a shorter length of stay (mean LOS 5.84) while old patients were likely to have relatively longer length of stay (mean LOS 11.77).

At level 3, for all but one group of young patients (young patients with diagnosis of other strokes), the covariate gender did not provide prognostically significant splits. While at level 3, for groups of old patients with stroke diagnosis hemorrhagic stroke, TIA and other strokes (node 7, node 11 and node 13) the covariate gender provided prognostically significant splits). For the group of old patients with cerebral infarction (node 9) the covariate gender split is not prognostically significant.

## III. The Extended Mixture Distribution Survival Tree Construction

This mixture distribution survival tree method can be extended to examine the relationship between the treatment outcome and patients' length of stay distribution and their interrelationship with patient characteristics by further partitioning each group of patients (determined using mixture distribution survival tree method above) into subgroups with more homogeneous patient pathways by covariate 'treatment outcome'. Although the information about the treatment outcome is not available at the time of admission, we can assign the probability to each treatment outcome using cohort analysis. The covariate ' treatment outcome ' can have any of the two values death or discharge from the hospital.

TABLE II
Extended Mixture Distribution Survival Tree Construction (Nodes And Possible Splits)

| Node | Covariate | Covariate value | Number of patients | Mean LoS | Standard deviation (LoS) | Coxian-phase type distribution | | | Gaussian mixture distribution | | | Improvement in AIC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Loglikelihood ($L_{max}$) | Number of phases | AIC | Loglikelihood | Number of phases | AIC | |
| **All** | **Complete dataset** | **Root node** | **1985** | **29.0106** | **52.8382** | **-8407.8007** | **3** | **16825.60** | **-8481.625** | 10 | **17021.25** | - |
| **Level 1** | | | | | | | | | | | | |
| 1 (Root node) | Gender | Male | 933 | 26.5938 | 44.0575 | -3859.8125 | 2 | 7725.625 | -3897.236 | 7 | 7834.47 | 10.2825 |
| | | Female | 1052 | 31.154 | 59.4698 | -4539.8469 | 3 | 9089.694 | -4569.59 | 9 | 9191.18 | |
| | Age | Young | 624 | 19.2564 | 39.1523 | -2316.974 | 2 | 4639.948 | -2342.454 | 7 | 4724.9 | 127.159 |
| | | Old | 1361 | 33.4827 | 57.4932 | -6024.2472 | 3 | 12058.49 | -6047.846 | 9 | 12147.7 | |
| | **Diagnosis** | **Hemorrhagic** | **154** | **33.6039** | **56.4456** | **-659.05019** | **3** | **1328.1** | **-665.5723** | **6** | **1365.14** | **306.1832** |
| | | **Cerebral** | **655** | **36.6611** | **47.6753** | **-2973.8941** | **4** | **5961.788** | **-2980.867** | **6** | **5995.73** | |
| | | **TIA** | **425** | **9.31294** | **19.9516** | **-1298.6262** | **2** | **2603.252** | **-1316.503** | **6** | **2667** | |
| | | **Other** | **751** | **32.5433** | **65.0453** | **-3310.1386** | **2** | **6626.277** | **-3297.443** | **8** | **6640.89** | |
| **Level 2** | | | | | | | | | | | | |
| 2 Hemorrhagic | Gender | Male | 80 | 28.2 | 52.09832 | -317.63202 | 4 | 649.2640 | -328.0819 | 4 | 678.1638 | 12.50299 |
| | | Female | 74 | 39.4459 | 60.254 | -328.16667 | 3 | 666.3333 | -329.4363 | 6 | 692.873 | |
| | **Age** | **Young** | **50** | **24.56** | **55.117** | **-173.39875** | **4** | **360.7975** | **-187.3508** | **3** | **390.702** | **15.65113** |
| | | **Old** | **104** | **37.9519** | **56.561** | **-468.82587** | **4** | **951.6517** | **-477.7462** | **4** | **977.492** | |
| 3 Cerebral | Gender | Male | 302 | 33.70860 | 49.8833 | -1334.898 | 4 | 2683.796 | -1339.464 | 5 | 2706.93 | -2.395836 |
| | | Female | 353 | 39.18697 | 45.5501 | -1635.194 | 3 | 3280.388 | -1639.346 | 5 | 3306.69 | |
| | **Age** | **Young** | **194** | **24.0670** | **42.4506** | **-785.3629** | **3** | **1580.726** | **-793.676** | **5** | **1615.35** | **35.8404** |
| | | **Old** | **461** | **41.961** | **48.787** | **-2173.9068** | **2** | **4353.814** | **-2158.611** | **5** | **4345.22** | |
| 4 TIA | Gender | Male | 207 | 8.7005 | 22.6817 | -607.9547 | 2 | 1221.909 | -637.5377 | 4 | 1297.08 | -1.20391 |
| | | Female | 218 | 9.8945 | 16.9366 | -686.27346 | 3 | 1382.547 | -690.5091 | 5 | 1409.02 | |
| | **Age** | **Young** | **176** | **5.83523** | **11.1641** | **-455.8639** | **2** | **917.7278** | **-465.3673** | **4** | **952.735** | **24.78144** |
| | | **Old** | **249** | **11.7711** | **24.0154** | **-827.3716** | **2** | **1660.743** | **-823.7194** | **8** | **1693.44** | |
| 5 Other strokes | Gender | Male | 344 | 30.7413 | 43.4091 | -1490.577 | 4 | 2995.154 | -1490.229 | 6 | 3014.46 | 8.886526 |
| | | Female | 407 | 34.0663 | 78.7981 | -1808.118 | 2 | 3622.237 | -1864.211 | 3 | 3744.42 | |
| | **Age** | **Young** | **204** | **24.9608** | **43.76126** | **-818.1347** | **4** | **1650.269** | **-819.1983** | **6** | **1672.4** | **36.23604** |
| | | **Old** | **547** | **35.3711** | **71.1697** | **-2466.8858** | **2** | **4939.772** | **-2543.835** | **3** | **5103.67** | |
| **Level 3** | | | | | | | | | | | | |
| 6 Hemorrhagic Young | Gender | Male | 29 | 30.5172 | 69.1114 | -108.83289 | 2 | 223.6658 | -105.682 | 4 | 233.364 | -9.21382 |
| | | Female | 21 | 16.3333 | 22.8126 | -70.172765 | 2 | 146.3455 | -74.17359 | 3 | 164.347 | |
| **7 Hemorrhagic Old** | **Gender** | **Male** | **51** | **26.8823** | **39.2027** | **-211.39224** | **4** | **436.7845** | **-211.0736** | **4** | **444.147** | **2.30463** |
| | | **Female** | **53** | **48.6038** | **67.5821** | **-253.2813** | **2** | **512.5626** | **-245.9886** | **7** | **531.977** | |
| 8 Cerebral Young | Gender | Male | 104 | 24.6731 | 49.2715 | -420.88798 | 2 | 847.776 | -427.2601 | 4 | 876.52 | -3.15333 |
| | | Female | 90 | 23.36667 | 32.9415 | -361.0516 | 4 | 736.1032 | -364.2716 | 4 | 750.543 | |
| 9 Cerebral Old | Gender | Male | 198 | 38.4545 | 49.6696 | -903.58419 | 4 | 1821.168 | -903.381 | 4 | 1828.76 | -0.62756 |
| | | Female | 263 | 44.6008 | 47.9429 | -1259.34 | 2 | 2524.681 | -1251.528 | 6 | 2537.06 | |
| 10 TIA Young | Gender | Male | 88 | 5.7386 | 11.3263 | -224.74588 | 2 | 455.4918 | -232.5108 | 3 | 481.022 | -5.52344 |
| | | Female | 88 | 5.9318 | 10.9988 | -230.8797 | 2 | 467.7595 | -231.1457 | 4 | 484.291 | |
| **11 TIA Old** | **Gender** | **Male** | **119** | **10.8908** | **28.0847** | **-377.7327** | **2** | **761.4654** | **-401.7215** | **3** | **819.443** | **2.615394** |
| | | **Female** | **130** | **12.5769** | **19.5270** | **-444.56883** | **4** | **903.1377** | **-437.3312** | **4** | **896.662** | |
| **12 Other strokes Young** | **Gender** | **Male** | **119** | **30.1092** | **52.7719** | **-493.3325** | **3** | **996.665** | **-486.2055** | **6** | **1006.41** | **0.265422** |
| | | **Female** | **85** | **17.7529** | **24.6624** | **-322.7928** | **3** | **655.5857** | **-318.6695** | **3** | **653.339** | |
| **13 Other strokes Old** | **Gender** | **Male** | **225** | **31.0756** | **37.52** | **-987.5257** | **4** | **1989.051** | **-991.5326** | **4** | **2005.07** | **3.802914** |
| | | **Female** | **322** | **38.3727** | **87.1713** | **-1470.4587** | **2** | **2946.917** | **-1517.217** | **3** | **3050.43** | |

Each terminal node of the survival tree of Figure 1 is further partitioned into daughter nodes by the covariate 'treatment outcome'. We grow the tree if the split maximizes node homogeneity by minimizing the value of AIC and if at a node, there is no split providing significant improvement in AIC, the node is designated as a terminal node.

Figure 2 is the schematic representation of the extended mixture distribution survival tree for the length of stay data on stroke patients from the Belfast City Hospital. The resulting tree now has 19 terminal nodes. Only two terminal nodes (node 9 and node 26) are approximated by GMD and all other terminal nodes are approximated by C-PhD. Table 2 lists terminal nodes of the survival tree of figure 1, and possible splits of these nodes by the covariate 'treatment outcome'. Bold faced splits were selected for splitting the parent node. Parent nodes are represented by pale blue rows with treatment outcome "all". The column P/G specifies which distribution among C-PhD and GMD provides better fit. The total improvement in AIC is 115.655 and the new AIC of all the terminal nodes is 16282.27.

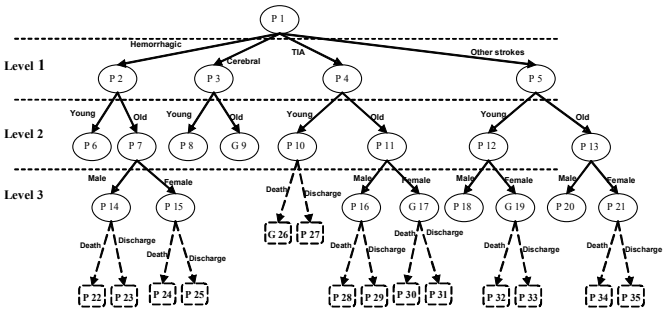


Fig. 2. Extended mixture distribution survival tree for the length of stay data on stroke patients from the Belfast City Hospital

## IV. Pathway Prognostication Using The Extended Mixture Distribution Survival Tree

The extended mixture distribution survival tree method can effectively be used to examine the relationship between LOS and treatment outcome at discharge and their interrelation with patient characteristics such as age, gender and diagnosis.

The extended mixture distribution survival tree clusters the length of stay data into 19 clinically meaningful patient groups, each representing a distinct patient pathway within the system. We can see that in seven patient groups (i.e., the terminal nodes in Figure 1), the covariate 'treatment outcome' has prognostic significance, i.e., patients with different treatment outcomes follow different patient pathways, while, there is homogeneity among patient pathways followed by the other five patient groups. The treatment outcome is prognostically most significant among the group of old male patients diagnosed with Hemorrhagic stroke. This also reflects with the difference in mean LOS. Among this group of patients, those are expected to discharge are likely to have longer length of stay (mean LOS 31.9) than those die in the course of their treatment (mean LOS 10.9). The Treatment outcome has prognostic significance for all groups of female patients while it has prognostic significance in only two groups of male patients (node 14 and node 16). All young patients with Hemorrhagic stroke followed homogeneous patient pathways. Treatment outcome does not have prognostic significance among all groups of patients diagnosed with cerebral infraction, while treatment outcome is prognostically significant in all groups of patients with TIA. Patients diagnosed with TIA and discharged from hospital are more likely to have shorter length of stay (mean LOS 5.24, 10.35, 11.16 respectively for node 27, node 29 and node 31) than those patients with TIA who died in the hospital (mean LOS 57.5, 21 and 48 respectively for node 26, node 28 and node 30).Prepare your paper in full-size format, on US letter paper 8 ½ by 11 inches). For A4 paper, use the A4 settings.

## V. Conclusion

Mixture distribution survival tree have advantage of achieving the improved within node homogeneity. Therefore, mixture distribution survival tree based analysis is a more effective method for prognostication of survival data and for clustering survival data into groups of patients following homogeneous patient pathways. It is a powerful method for determining the relationship between input covariates and outcome measures and their interrelations. It provides better understanding of heterogeneity of patient pathways stratified by covariates representing patient characteristics such as age, gender, diagnosis and outcome measures such as treatment outcome, destination at discharge. We can also use the model to estimate the length of stay of a patient based on his/her characteristics (age, gender, diagnosis) available at the time of admission. We can extend this approach by further growing the tree by partitioning the terminal nodes into subgroups with more homogeneous patient pathways based on covariates representing outcome measures. Although the information about the treatment outcome is not available at the time of admission, we can assign the probability to each treatment outcome using cohort analysis. This information can be used for estimating bed requirements for each group of patients (following homogeneous patient pathways) and capacity planning for the whole care system. As future work we will also assess the use of other mixture distributions in order to achieve further improvement in within node homogeneity. Presently we are developing application of our model for capacity planning in a stroke care unit.

## Acknowledgment

The authors acknowledge support for this work from the EPSRC (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

[1] L. Garg, S. I. McClean, B. J. Meenan, E. El-Darzi and P. H. Millard, "Clustering patient length of stay using mixtures of Gaussian models

and phase type distributions". *The 22nd IEEE Symposium on Computer-Based Medical Systems (CBMS 2009)*, Albuquerque, New Mexico, USA, August 3-4, 2009, pp. 1-7, doi:10.1109/CBMS.2009.5255245.

[2] R. Davis and J. Anderson., "Exponential Survival Trees", *Statistics in Medicine*, 8, 1989, pp. 947-962 .

[3] F. Gao, A. K. Manatunga and S. Chen, "Identification of prognostic factors with multivariate survival data", *Computational Statistics & Data Analysis*, 45, 2004, pp. 813-824.

[4] M. Barton, S. I. McClean, L. Garg and K. Fullerton, "Modelling Stroke Patient Pathways using Survival Analysis and Simulation Modelling", Eds: Leonidas Sakalauskas, Christos Skiadas, Edmundas K. Zavadskas, *Proceedings of the XIII International Conference on Applied Stochastic Models and Data Analysis (ASMDA 2009)*, 2009, pp. 370-373, ISBN: 978-9955-28-463-5, Publisher: *Vilnius Gediminas Technical University Press*.

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

[6] World Health Organisation, "International Statistical Classification of Diseases and Related Health Problems", Tenth Revision – ICD-10, second edition, *WHO*: Geneva , 2007.