CrossMark

Predicting elderly patient length of stay in hospital and community care using a series of conditional Coxian phase-type distributions, further conditioned on a survival tree

Andrew S. Gordon[1] · Adele H. Marshall[1] · Mariangela Zenga[2]

Abstract Increasing demand on hospital resources by an ageing population is impacting significantly on the number of beds available and, in turn, the length of time that elderly patients must wait for a bed before being admitted to hospital. This research presents a new methodology that models patient pathways and allows the accurate prediction of patient length of stay in hospital, using a phase-type survival tree to cluster patients based on their covariates and length of stay in hospital. A type of Markov model, called the conditional Coxian phase-type distribution is then implemented, with the probability density function for the time spent at a particular stage of care, for example, the first community discharge, conditioned on the length of stay experienced at the previous stage, namely the initial hospital admission. This component of the methodology is subsequently applied to each cohort of patients over a number of hospital and community stages in order to build up the profile of patient readmissions and associated timescales for each cohort. It is then possible to invert the methodology, so that the length of stay for an observation representing a new patient admission may be estimated at each stage of care, based on the assigned cohort at the initial hospital stage. This approach provides hospital managers with an accurate understanding of the rates with which different groups of patients move between hospital and community care, which may be used to reduce the negative effects of bed-blocking and the premature discharge of patients without a required period of convalescence. This has the benefit of assisting hospital managers with the effective allocation of vital healthcare resources. The approach presented is different to previous research in that it allows the inclusion of patient covariate information into the methodology describing patient transitions between hospital and community care stages in an aggregate Markov process. A data set containing hospital readmission data for elderly patients from the Abruzzo region of Italy is used as a case study in the application of the presented methodology.

Keywords Readmissions · Survival tree · Coxian phase-type distribution · Hospital predictions · Length of stay

1 Introduction

National health care systems have a limited amount of resources heavily in demand which have to be carefully managed in order to offer efficient hospital care for their population. Elderly patient care accounts for the greatest use of hospital resources [1], due to issues surrounding frailty and an increased number of comorbidities contributing to hospital readmissions. However, in addition to this, it is often the case that elderly patients must remain in hospital after they are declared fit to be discharged, because they are waiting for an available place in community care. In the United Kingdom National Health Service, this is known as bed-blocking and represents an enormous consumption of hospital resources that could be avoided if better resource planning was in place [2]. Advanced knowledge, in the

✉ Andrew S. Gordon
agordon17@qub.ac.uk

1   Mathematical Sciences Research Centre,
    Queen's University Belfast, University Road, Belfast,
    BT7 1NN, UK

2   Department of Statistics and Quantitative Methods, University
    of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126
    Milano, Italy

form of accurate predictions, for when elderly patients are expected to leave hospital has the potential to help secure, in advance, the community care anticipated, thereby reducing or even eliminating this waste of vital resources. This is particularly important in ensuring that the proper care resources are available for elderly patients readmitting to hospital after a number of spells in both hospital and community care.

Previous research has shown that statistical models can accurately capture the time taken for the progression of patients through care as a single stage [3, 4]. The focus of the current research is to describe the movement of elderly patients through multiple ordered stages, from the initial hospital admission, to the first community stage, then the first hospital readmission stage, followed by the second community stage and finally the second hospital readmission stage. Nevertheless, due to the inherent variability in the time spent at each stage in the pathway for elderly patients, an approach which makes predictions based on the assumption that the elderly population is homogeneous with respect to length of stay, is often inaccurate [5]. Although length of stay may appear a simple metric, it can be influenced by a number of factors including reason for admission, hospital policy and issues extending beyond the hospital environment.

This paper introduces a methodology which can account for some of this variability by making predictions for subgroups of the population, known as cohorts, each of which may be modelled by a separate distribution. This is carried out by extending the conditional Coxian phase-type distribution [6], a method which calculates the conditional probability for length of stay at the current stage of care, based on the length of stay experienced at the previous care stage, to be further conditioned on a phase-type survival tree [7]. This latter technique may be used to partition elderly patients into cohorts based on their length of stay at the initial hospital stage, in such a way so that patients in different cohorts have a significantly different length of stay distribution. This allows the conditional Coxian phase-type distribution to be subsequently employed to model the survival data in each cohort through until the second hospital readmission stage, taking into account the length of stay at the previous stage of care, for each cohort separately. The research presented allows the identification of multiple patient care pathways, arising through specific combinations of elderly patient covariates, with the illustration of the aforementioned covariates provided in a tree-like structure. The resulting parameter estimates from the implementation of the conditional Coxian phase-type distribution to model the pathway for each cohort, may be used to predict the length of stay for an elderly patient with a given combination of covariates at a particular stage of care, thereby allowing health care planners to more effectively allocate vital health care resources.

The remainder of this paper is structured as follows: Section 2 introduces the methodology implemented in the current research, including the phase-type survival tree and the conditional Coxian phase-type distribution process component. In Section 3, an overview of the data set containing hospital readmission data for elderly patients admitted to hospitals in the Abruzzo region of Italy is presented. This section also contains the application and validation of the proposed methodology, along with prediction of length of stay for all of the identified cohorts throughout each stage of care. Conclusions, together with a discussion on the points raised from the implementation of the research are presented in Section 4.

2 Methodology

2.1 Coxian phase-type distribution

Phase-type distributions may be used to describe the time until absorption of a finite Markov chain in continuous time, where there is a single absorbing state and the process begins in a transient state [8]. Coxian phase-type distributions [9] incorporate a series of transient states in order to describe the duration until an event of interest occurs, where the system begins in the first transient state and proceeds in a sequential fashion through the transient states. The system may not move more than one transient state at a time, neither is it permitted to move backwards to any previous state. The method determines the rates associated with the movement of the system between the latent states, before absorption occurs, representing the occurrence of the event. In this respect, the Coxian phase-type distribution is particularly useful for modelling elderly patient movements through hospital [10–12], where the event of interest is the patient leaving hospital [13]. In this research, elderly patients may leave hospital through one of two different scenarios: death or discharge to the community. This gives rise to the inclusion of two absorption states in the Coxian phase-type distribution. More formally, let $X(t); t \geq 0$ be a Markov chain in continuous time with states $1, 2, ..., m, m + 1, m + 2$, where states $m + 1$ and $m + 2$ are the two absorbing states, the rest being transient and ordered. With the system beginning in state 1: $X(0) = 1$, Fig. 1 shows a representation of the Coxian phase-type distribution with two absorbing states, where transitions occur in a small time interval, h. The transition rate from transient state i to the next transient state, $i + 1$, is denoted by λ_i and the transition rate from transient state i to absorbing state k is denoted by μ_i^k.

This approach uses available prior information in the form of an indicator variable, describing which of the two events has occurred for each observation. This is to make sure that only the absorption rates for that mode of absorption
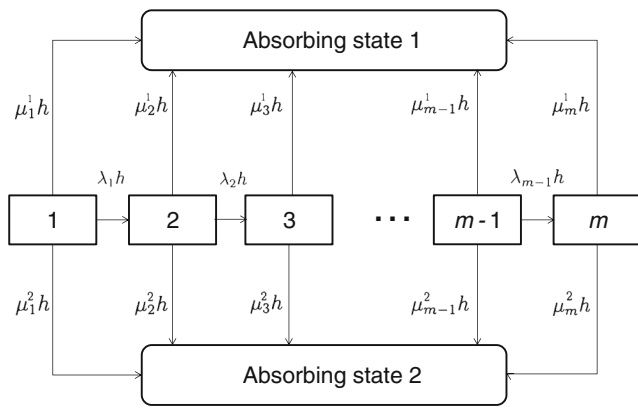
Fig. 1 Coxian phase-type distribution with m transient states and two absorbing states

are affected by instances of the system having that particular event. The generator matrix $\boldsymbol{Q}$, characterising the Markov process, for the distribution of times until event k occurs is given by

$$
\boldsymbol{Q} = \begin{pmatrix}
-(\lambda_1 + \mu_1^k) & \lambda_1 & 0 & \ldots & 0 & 0 \\
0 & -(\lambda_2 + \mu_2^k) & \lambda_2 & \ldots & 0 & 0 \\
0 & 0 & -(\lambda_3 + \mu_3^k) & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & -(\lambda_{m-1} + \mu_{m-1}^k) & \lambda_{m-1} \\
0 & 0 & 0 & \ldots & 0 & -\mu_m^k
\end{pmatrix}
\tag{1}
$$

The probability density function for the Coxian phase-type distribution illustrated in Fig. 1 is given by

$$
f(t) = \boldsymbol{p} \exp(\boldsymbol{Q}t)\boldsymbol{q} \tag{2}
$$

where

$$
\boldsymbol{p} = (1, 0, 0, ..., 0) \tag{3}
$$
$$
\boldsymbol{q} = (\mu_1^k, \mu_2^k, ..., \mu_m^k)^T \tag{4}
$$

It should be noted that the transition rates between transient states are common to the processes leading to the occurrence of both events and are not superscripted with k for this reason.

2.2 Conditional Coxian phase-type distribution

The Coxian phase-type distribution may be used to model the movement of elderly patients through the initial hospital stage, with two possible eventual outcomes: death or discharge to the community. However, with the aim of this research to model the movement of elderly patients through an ordered sequence of care stages, an approach is employed which can take into account the length of stay at a previous stage in the determination of transition rates for the current stage of care. Such an approach is the *conditional Coxian phase-type (CCPh) distribution* [6] where the system of stages is considered using two stages at a time. Once

the information from the first stage is used to inform the distributional form for the second stage, the process repeats, whereby the second stage is used to inform that for the third stage, and so on. This is so as to include as much information as possible in the determination of transition rates for each stage of care. This is achieved through the use of Bayes' theorem in the probability density function, where t_1 is the length of time spent at the previous care stage, denoted by A and t_2 is the length of time spent at the current care stage, denoted by B:

$$
P(B = t_2 | A = t_1) = \frac{P(A = t_1 \cap B = t_2)}{P(A = t_1)} \tag{5}
$$
$$
= \frac{\boldsymbol{p}_A \exp(\boldsymbol{Q}_A t_1) \mathbf{T}_{AB} \exp(\boldsymbol{Q}_B^k t_2) \boldsymbol{q}_B^k}{\boldsymbol{p}_A \exp(\boldsymbol{Q}_A t_1) \boldsymbol{q}_A} \tag{6}
$$

where the notation is as described in Section 2.1, subscripted with A and B to reflect the previous and current stages respectively. Each element $\boldsymbol{T}_{AB}(i, j) = \mu_{A_i B_j}$, represents the rate of transition between the ith state of stage A and the jth state of stage B. However, to meet the requirement that patients may only enter stage B at the first state, $\boldsymbol{T}_{AB}$ contains non-zero rates in the first column only. As such, it takes the following form, where k_a is the number of transient states in stage A.

$$
\mathbf{T}_{AB} = \begin{pmatrix}
\mu_{A_1 B_1} & 0 & \ldots & 0 \\
\mu_{A_2 B_1} & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\mu_{A_{k_a} B_1} & 0 & \ldots & 0
\end{pmatrix}
\tag{7}
$$

Equation 6 may be used as the probability density function for the conditional Coxian phase-type distribution with two absorbing states. It is important to note that both the optimal parameter estimates from the implementation of the methodology for the previous stage of care and the times experienced at the previous stage of care are necessary in Eq. 6. Additionally, there are no superscript k values in Eq. 6 as those patients who have entered stage B, from stage A, have done so through the absorbing state representing community discharge in stage A. The remaining patients have already left the aggregate system through the alternative absorbing state in stage A (representing death) and therefore are not considered for stage B.

2.3 Phase-type survival tree

The methodology described so far has been concerned with the distributional fit for a sequence of skewed survival distributions. However, with the application under consideration that of elderly patient care, there is often a large amount of variability in length of stay in care, due to the wide-ranging circumstances surrounding admission and discharge from hospital. As a result, elderly patient length of stay is usually not homogeneous, and a technique is sought which

can account for a large proportion of this variability. Mainstream data mining techniques have been considered, for example, logistic regression, decision trees, Bayesian networks and random forests. However such methods require the presence of a categorical response variable, the levels of which both have significantly different distributions of the survival variable and may be accurately predicted using the remaining variables in the data set. Given the application under consideration, it is often difficult to identify a suitable candidate response variable, due to the complicated set of circumstances usually surrounding each individual elderly patient admission to hospital. The destination of the patient on leaving hospital, for example, death or discharge, is often associated with significantly different length of stay distributions. However, this detail is accounted for through the provision of separate absorbing states in the component of the methodology which models the survival distributions at each stage of care.

With the above considerations in mind, an alternative approach is presented, which does not require the partitioning of patient observations at the indication of a single variable. Instead, observations may be separated into subgroups based on whether or not splitting using a sequence of variables results in a significant reduction in the overall variability of the data. A technique which intrinsically partitions observations into subgroups is a phase-type survival tree [7, 14, 15]. This is carried out through splitting observations into cohorts based on their characteristics (or covariates), so that patients in the same cohort have a similar distribution for length of stay and patients in different cohorts have significantly different distributions for length of stay. Using a survival tree, the heterogeneity apparent in the data is reduced to multiple smaller sets of data which are each homogeneous with respect to a particular survival distribution. The survival tree is constructed using Coxian phase-type distributions, with the Akaike information criterion (AIC) [16] corresponding to the optimal distribution used as the splitting criterion. For this reason, the method is called the *phase-type* survival tree. This component of the methodology is applied to observations representing elderly patient admissions at the initial hospital stage only.

Figure 2 shows a simple conceptual diagram of how the survival tree and conditional Coxian phase-type distribution components are used in conjunction. All of the elderly patient records referring to an initial hospital admission are used as input to determine the root node variable of the survival tree, denoted by variable A. To determine the root node, the Coxian phase-type distribution is fitted to all observation times, with the optimal AIC value recorded. The data set is then partitioned according to both variable and variable level. The Coxian phase-type distribution is fitted to the observations within each level, with the resulting optimal AIC values summed to obtain a total AIC value

for each variable. The variable corresponding to the greatest improvement in AIC value from that corresponding to the un-partitioned fit is selected as the root node. Once this variable has been identified, the records are split according to the different levels of variable A with the same process used once again to determine variables B and C from the remaining pool of variables. This process is repeated until the survival tree is constructed, at which point the Coxian phase-type distribution (with two absorbing states) fitting each leaf of the tree represents the initial hospital length of stay data for the respective patient cohort. In the example given by Fig. 2, there are four identified patient cohorts, numbered '1' through to '4'. The methodology then considers the next stage of the aggregate process, the first community stage. The conditional Coxian phase-type distribution is fitted to each cohort separately for those patients undergoing this stage and the rates of transition for this stage of care determined. This is carried out for the remaining stages of care in the aggregate process (pathway) until the second hospital readmission stage. Once placed into cohorts based on the initial hospital stage data, patients remain in these cohorts until either the time-frame for the study ends (the second hospital readmission stage), or they leave one of the care stages through death or they do not require a further readmission to hospital.

3 Application

3.1 Tree-building process and identification of cohorts

The methodology presented in this research is applied to a data set consisting of hospital readmission information for elderly patients in the Abruzzo region of Italy. To the best of the authors' knowledge, the hospitals from which the data is collected provide very similar levels of care to elderly patients, with rehabilitation taking place at additional facilities, for example, specialised hospitals and residential care. As such, each of the hospitals may be assumed to provide adequate care for all elderly patients in the data set. The data set contains length of stay times for 7,251 patients, of which 1,067 experience a readmission and 261 patients undergo a further readmission, between the 10th November 2008 and the 31st of December 2009. Due to the relatively short duration of the study and in keeping with previous literature [17] incorporating Italian patient readmissions, all subsequent admissions to hospital for a particular patient, within this time interval, are classified as a readmission. The purpose of the study is to provide a methodology which is capable of modelling the movement of elderly patients between hospital and community care, with a view to addressing the clinical need for a more effective allocation of healthcare resources. As such, the data set under consideration

Fig. 2 Conceptual schema of the approach with the survival tree identifying cohorts based on the initial hospital stage before the conditional Coxian phase-type distribution is applied successively to each cohort
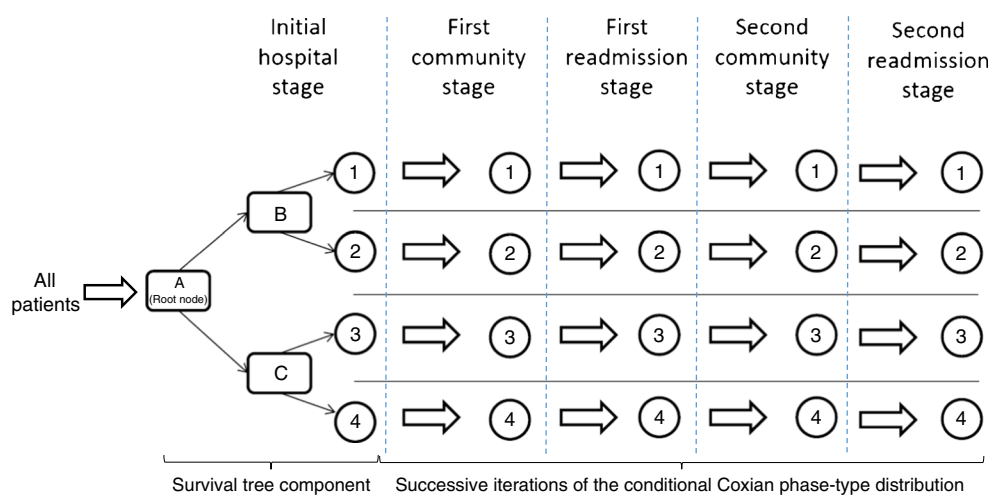
does not include information on the specific type of care that patients receive in or between hospital spells, only whether they have left each hospital stage through death or discharge. Furthermore, the data set contains information on only those patients who are readmitted to hospital from the community; there is no information on patients who have died in the community, or indeed, those who are still alive in the community without requiring an additional readmission to hospital. However, with the focus of the current research to provide an insight into the readmission patterns of elderly patients in order to facilitate hospital managers in the allocation of hospital resources, data on the patients who are readmitted to hospital is what is essential to do so. Consequently, when employing the methodology to model community stages, only one absorbing state is used, representing readmission to hospital.

Patients in the study are aged between 65 and 104 years, with lengths of stay ranging from 1 to 85 days in hospital care. There are three variables of interest, representing patient covariates, recorded in the data set: age, gender and admission method. The age variable has been discretised to take values < 76, $76-82$ and > 82 years respectively. These intervals are chosen as they simultaneously indicate significantly different distributions of length of stay for patients between the levels (p-value = 0.0289) so as to facilitate the construction of the survival tree, whilst also containing large enough numbers of observations in each level so as to eliminate any class imbalance. This latter reason is particularly important because many patient observations are lost from each cohort as the number of readmissions increases, meaning that at the second hospital readmission stage there is a risk of having insufficient observations over which to run the methodology. To this end, it is desirable to have roughly representative numbers of patients at the each of the hospital stages to demonstrate the working of the methodology, given the relatively small data set. Other age intervals, for example age bands of ten years, have been considered in this

research, however this results in a greater number of intervals leading to smaller numbers of observations within each cohort. Additionally in other tested partitions, the lengths of stay for the variable levels are either not significantly different, or result in the introduction of class imbalance.

The admission method variable also consists of three levels: 'Emergency', 'Planned' and 'Other', giving an indication as to the urgency of the admission to hospital for each patient. The vast majority of hospital admissions are recorded as either 'Emergency' or 'Planned', with the small proportion ($<3\%$) representing 'Other' perhaps arising from rare instances which do not fall broadly within either of the aforementioned categories, or have simply been unrecorded. They have been included in this analysis due to the results of running the methodology over a similar additional data set of elderly patient records from an Italian region, where elderly patients having an admission method categorised as 'Other' were found to have significantly different lengths of stay from those categorised as either 'Emergency' or 'Other', thereby indicating a third admission method in its own right. Table 1 details the construction of the survival tree, where the variable showing the greatest AIC improvement is chosen to partition at that point in the construction of the tree.

Application of the phase-type survival tree to the 7,251 observations from the initial hospital stage results in the identification of seven cohorts, with the survival tree shown in Fig. 3. The AIC value for 7,251 observations without using any variables to split is 44,508. In the determination of the root node for the survival tree, only the admission method variable is successful in reducing the variability in the length of stay variable (AIC improvement of 3). Therefore, this variable is used for splitting at the root node. The resulting branches representing patients entering hospital as 'Emergency' and 'Planned' admissions show additional improvement when split by both Age and Gender, with Age in both cases representing optimal improvement (AIC

Table 1 Numerical construction of the survival tree

Node	Variable	Variable level	Patients	*L*	No. of states	AIC	Improvement
All	All	All	7251	−22237	6	44508	−
1	Age	< 76	910	−2831	4	5683	
Root node		76-82	2679	−8134	5	16297	-4
		>82	2535	−6233	5	12480	
	Admission	Emergency	4995	−15300	5	30628	
	method	Planned	2052	−6325	5	12678	+3
		Other	204	−592	3	1199	
	Gender	Female	4058	−9584	6	24996	
		Male	3193	−9745	5	19517	−5
2	Age	< 76	430	−1356	3	2728	
Admission		76-82	1846	−5589	7	11198	+27
method		> 82	2719	−8320	6	16674	
(Emergency)	Gender	Female	2802	−8567	6	17167	
		Male	2193	−6705	7	13449	+11
3	Age	< 76	471	−1289	4	2601	
Admission		76-82	762	−2311	5	4650	+287
method		> 82	819	−2556	5	5140	
(Planned)	Gender	Female	1139	−3530	4	7082	
		Male	913	−2760	5	5548	+48
4	Age	< 76	9	−17	2	40	
Admission		76-82	71	−210	3	437	−3
method		> 82	124	−355	3	725	
(Other)	Gender	Female	117	−342	3	701	
		Male	87	−250	3	515	−17
5	Gender	Female	200	−617	3	1251	
Adm (Emerg.)		Male	230	−739	3	1494	−17
Age ¡76							
6	Gender	Female	911	−2758	6	5549	
Adm (Emerg.)		Male	935	−2831	6	5696	−46
Age 76-82							
7	Gender	Female	1691	−5187	5	10401	
Adm (Emerg.)		Male	1028	−3130	6	6294	−21
Age ¿82							
8	Gender	Female	203	−644	3	1303	
Adm (Planned)		Male	268	−785	5	1597	−299
Age ¡76							
9	Gender	Female	409	−1246	4	2513	
Adm (Planned)		Male	353	−1072	4	2166	−29
Age 76-82							
10	Gender	Female	527	−1662	5	3351	
Adm (Planned)		Male	292	−889	4	1800	−11
Age ¿82							

Fig. 3 Survival tree to identify cohorts based on the initial hospital stage
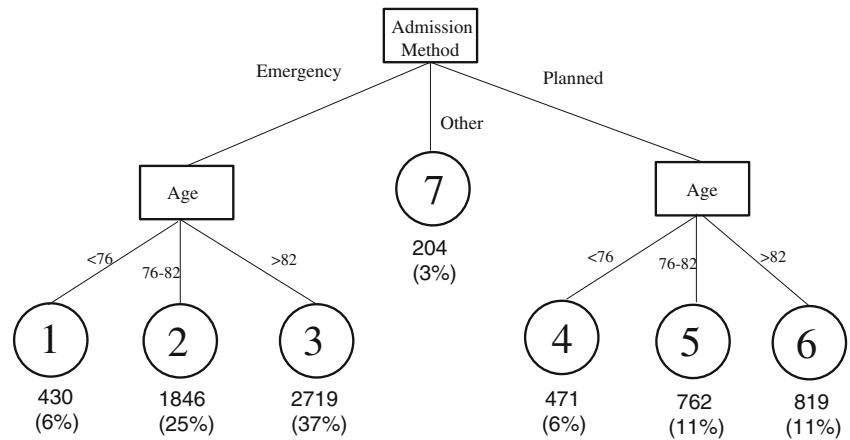

improvements of 27 and 287 respectively). The node resulting from the 'Other' admission method level becomes a leaf, due to the absence of significant improvement when splitting using age or gender. For the remaining nodes, splitting using the only other available variable, gender, does not result in a significant reduction of variability in the length of stay variable. The fitted distributions for both death and discharge patients are plotted in Fig. 4.

Figure 4 shows that the survival tree has been successful in identifying cohorts of patients with significantly different distributions for length of stay. This is particularly true for the case where patients leave hospital through death, with the survival curves visibly disparate. However, the benefit of using the survival tree may also be seen when patients who are discharged from hospital are considered, although

they are less visibly distinct from one another. Nevertheless, although apparently similar, the Coxian phase-type distributions representing the seven cohorts generally have differing numbers of phases, ranging from three to seven, thereby supplying additional evidence that the distribution of times between the cohorts are significantly different.

As patients move into the subsequent stages of care, they remain in the cohorts to which they were assigned at the initial hospital stage. Table 2 shows the number of elderly patients at each stage of care by cohort. It should once again be noted that when elderly patients leave a particular hospital spell, they may do so through either death or community discharge. Those patients who are discharged may die in the community, remain alive whilst not requiring a further hospital readmission or be readmitted to hospital. It is only
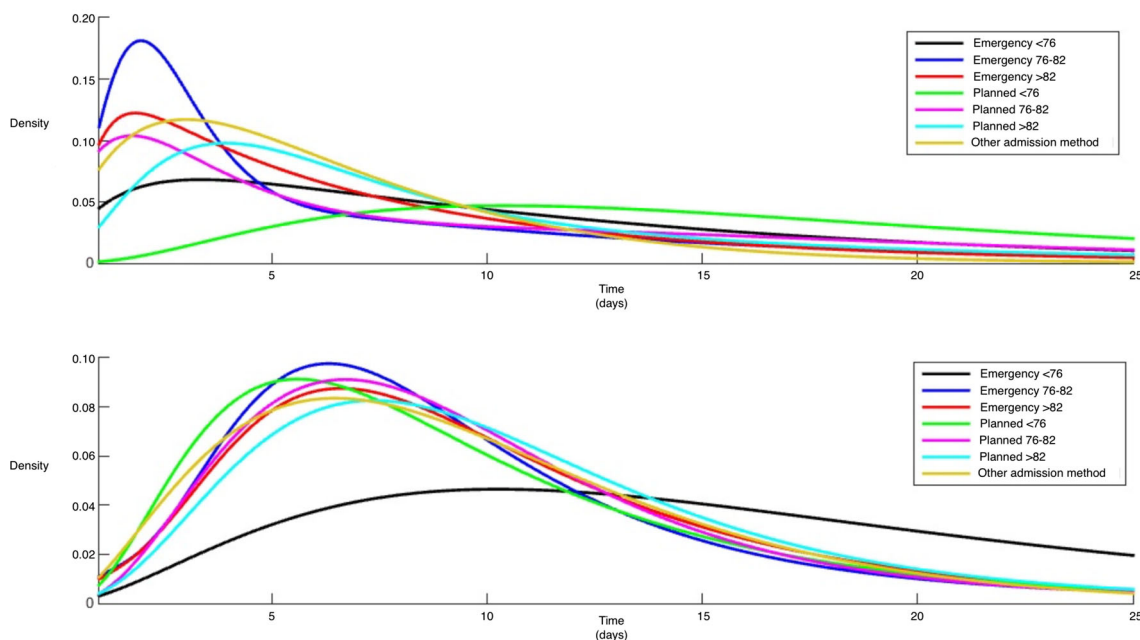
Fig. 4 Distributional fits to the cohorts in the initial hospital stage for death (*above*) and discharge (*below*)

Table 2 Number of patients in each admission for every cohort

Cohort	Initial admission	First readmission	Second readmission
Emergency; < 76	430	56	18
Emergency; $76 - 82$	1846	272	68
Emergency; > 82	2719	439	111
Planned; < 76	471	42	6
Planned; $76 - 82$	762	103	23
Planned; > 82	819	121	23
Other	204	34	12

this latter group of patients which are accounted for in the data set, explaining why the numbers between successive hospital readmissions decrease by such a high factor. Additionally, this also means that the number of patients in each community stage is the same as the number of patients in the subsequent hospital readmission stage. After the initial hospital stage, each community and hospital readmission stage is represented using a conditional Coxian phase-type distribution with one and two absorbing states respectively.

3.2 Validation of methodology and prediction of length of stay for new observations

The above methodology may be validated through the comparison of the empirical median length of stay value for a given cohort of patients at a particular stage, with the median value of simulated data obtained through the use of the associated fitted distribution. The median is chosen due to the high level of positive skewness in the data. Specifically, a set of times for each fitted distribution is simulated and the median value is calculated. This fitted median value is compared with the 95% confidence interval for the median value of empirical times for the corresponding cohort and stage. If the fitted median value lies within this confidence interval, then that distribution is said to be representative of the underlying data used to fit it. The methodology as a whole is validated if this is true for a high proportion of the fits for each cohort and stage combination. A 95% confidence interval for the median value of a distribution may be obtained by using Eq. 8 to calculate the Normal approximation to the Binomial distribution:

$$(t_{[np-1.96\sqrt{np(1-p)}]}, t_{[np+1.96\sqrt{np(1-p)}]}) \qquad (8)$$

where $t_{[i]}$ is the ith ordered time, n is the number of observations and $p = 0.5$ to specify that the confidence interval is calculated for the median. The simulation of times for a given distribution may be carried out through the use of the survivor function for that distribution. For a phase-type distribution, the survivor function is given by:

$$S(t) = \boldsymbol{p} \exp(\boldsymbol{Q}t)\boldsymbol{e} \qquad (9)$$

where $\boldsymbol{e}$ is the unity column vector. A simulated time from this distribution may be obtained by first selecting a random number from the uniform distribution, substituting this value for $S(t)$ and using the Newton-Raphson method [18] to solve for t. This process is repeated until n times are simulated. It should be noted that this validation step is carried out to ensure that the individual conditional Coxian phase-type distribution components have each captured the time-distributional features of the underlying data. Once verified, focus then moves to the determination of predictive intervals for each cohort/stage combination using a slightly different approach.

3.2.1 Validation of methodology for the Abruzzo data set

The distributions for patients leaving hospital through death and discharge are validated separately, since each hospital stage is represented by two processes; for patients who leave through death and discharge. This gives rise to an extra three distributions (since there are three hospital admissions considered: the initial admission, first readmission and second readmission) for each of the seven cohorts, making a total of 56 fitted distributions to be validated. Only three fitted median values do not lie within the 95% confidence interval for the empirically calculated medians, giving an accuracy of 94.6%. The three distributions which do not validate the methodology are all in the second hospital readmission stage for patients leaving through death, and as such, have an extremely small number of observations in the distribution ($n = 1$, 1 and 2). In each case and because of the lack of data for this distribution, the empirical 95% confidence interval for the median value is constrained to have a range equal to a single value, for example, (2, 2), meaning that for a model using continuous data, it is very unlikely for the representative value to fall within this interval. In any case, the authors are confident that this is an issue simply with the quantity of data available for later stages of care and that the survival tree has been successful in accounting for the variation in length of stay across all considered stages of care, even as far along as the second hospital readmission in most cases. The fitted median value and 95% confidence interval for each cohort at every stage are shown in Table 3,

Table 3 Validation of the methodology and predicted intervals for each cohort

Stage	Number of patients	Fitted median	Empirical 95% CI	Predicted 95% CI
Cohort 1 - Emergency; Age < 76				
H_0 (Death)	35	9	(5, 13)	(8, 9)
H_0 (Discharge)	395	9	(9, 10)	(9, 10)
C_1	56	47	(32, 72)	(42, 50)
H_1 (Death)	4	8	(4, 14)	(7, 8)
H_1 (Discharge)	52	8	(7, 9)	(7, 8)
C_2	18	31	(8, 70)	(28, 33)
H_2 (Death)	1*	3	(4, 4)	(3, 3)
H_2 (Discharge)	17	7	(4, 13)	(6, 8)
Cohort 2 - Emergency; Age $76 - 82$				
H_0 (Death)	177	4	(3, 6)	(4, 5)
H_0 (Discharge)	1669	8	(8, 8)	(8, 9)
C_1	272	54	(41, 59)	(51, 59)
H_1 (Death)	41	4	(2, 6)	(4, 5)
H_1 (Discharge)	231	8	(7, 9)	(8, 9)
C_2	68	38	(28, 52)	(35, 42)
H_2 (Death)	4	4	(1, 9)	(4, 5)
H_2 (Discharge)	64	8	(6, 10)	(8, 9)
Cohort 3 - Emergency; Age > 82				
H_0 (Death)	399	6	(5, 6)	(5, 6)
H_0 (Discharge)	2320	9	(8, 9)	(8, 9)
C_1	439	56	(50, 64)	(50, 63)
H_1 (Death)	79	5	(3, 7)	(4, 5)
H_1 (Discharge)	360	8	(8, 9)	(8, 9)
C_2	111	42	(25, 51)	(39, 46)
H_2 (Death)	14	4	(2, 6)	(4, 4)
H_2 (Discharge)	97	7	(6, 7)	(6, 7)
Cohort 4 - Planned; Age < 76				
H_0 (Death)	14	14	(9, 18)	(13, 15)
H_0 (Discharge)	457	8	(8, 9)	(8, 8)
C_1	42	69	(53, 109)	(63, 76)
H_1 (Death)	3	7	(2, 16)	(7, 8)
H_1 (Discharge)	39	8	(7, 11)	(8, 9)
C_2	6	44	(1, 73)	(40, 48)
H_2 (Death)	1*	4	(6, 6)	(4, 4)
H_2 (Discharge)	5	5	(3, 18)	(4, 5)
Cohort 5 - Planned; Age $76 - 82$				
H_0 (Death)	36	7	(3, 13)	(6, 7)
H_0 (Discharge)	726	9	(8, 9)	(9, 9)
C_1	103	64	(59, 92)	(60, 71)
H_1 (Death)	7	4	(1, 7)	(4, 4)
H_1 (Discharge)	96	9	(8, 10)	(8, 9)
C_2	23	56	(32, 96)	(52, 61)
H_2 (Death)	4	6	(5, 16)	(6, 7)
H_2 (Discharge)	19	6	(3, 14)	(5, 6)
Cohort 6 - Planned; Age > 82				
H_0 (Death)	72	7	(6, 8)	(7, 7)
H_0 (Discharge)	747	10	(9, 10)	(9, 10)
C_1	121	63	(44, 77)	(57, 68)

Table 3 (continued)

Stage	Number of patients	Fitted median	Empirical 95% CI	Predicted 95% CI
H_1 (Death)	16	4	(2, 7)	(4, 4)
H_1 (Discharge)	105	9	(8, 10)	(9, 10)
C_2	23	41	(29, 52)	(39, 44)
H_2 (Death)	6	5	(2, 7)	(4, 5)
H_2 (Discharge)	17	8	(3, 14)	(7, 8)
Cohort 7 - Other admission method				
H_0 (Death)	29	5	(4, 7)	(5, 5)
H_0 (Discharge)	175	9	(8, 10)	(8, 9)
C_1	34	40	(12, 86)	(35, 48)
H_1 (Death)	3	2	(1, 3)	(2, 2)
H_1 (Discharge)	31	7	(6, 9)	(7, 8)
C_2	12	33	(4, 84)	(31, 36)
H_2 (Death)	2*	1	(2, 2)	(1, 1)
H_2 (Discharge)	10	6	(4, 12)	(6, 7)

where the care stages have been abbreviated to H_i and C_j, denoting the ith hospital stage ($i = 0, 1, 2$ denoting the initial admission, first readmission and second readmission) and j^{th} community stage ($j = 1, 2$ denoting first and second discharge), respectively. The three instances where the fitted median does not fall within the empirical confidence interval have the number of observations marked with a *.

3.2.2 Prediction of length of stay for new observations for the Abruzzo data set

Having used simulations to validate the methodology, a similar approach may be used to predict the length of stay for a new elderly patient arriving to hospital, in addition to predictions for length of stay in potential successive stages of care. Upon arrival to hospital for their initial admission, the survival tree may be used to classify the patient into a cohort based on the method of admission and their age (both of which would be available at the point of entering the hospital). Confidence intervals for the median values of the simulated sets of times may then be calculated, through the use of Eq. 8 and the fitted distributional form for each cohort, serving as predictions on when the patient is likely to leave the stages of care they enter, beginning with the initial hospital stage and extending through to the second hospital readmission stage.

Table 3 shows the prediction intervals for each of the seven cohorts across all stages of care. These results indicate that elderly patients generally spend less time in hospital, across all of the cohorts, as the number of readmissions increase. This may be because hospital staff better know the medical requirements of their patients, with increased hospital readmissions, although this is assuming that patients are

readmitted to the same hospital department on their readmission. Furthermore, elderly patients who are admitted as planned admissions generally stay for a longer period in hospital than those who are admitted as emergency cases. Upon consultation with a clinical director and geriatrician of significant experience in this field, this is once again not an unreasonable result. This may be because emergency cases are generally regarded as more urgent than planned admissions and the case may be that they are treated as more of a priority, thereby leading to a shorter duration in hospital. This often results in planned admissions waiting longer for the treatment they require. This trend observed from the methodology output agrees with that obtained by simple calculation of the median length of stay for elderly patients as a whole, which shows a decrease from 8 days at the initial hospital stage to 7 days at the second hospital readmission stage. The methodology presented in this research paper has since been applied to a set of elderly patient records from hospitals in a different region, with the same general trends evident.

4 Conclusion and Discussion

The research presented in this paper extends previous methodologies which aim to model the pathway of elderly patients between various types of care [6, 19] to a methodology which incorporates patient characteristics, in order to provide predictions on length of stay for cohorts of elderly patients having significantly different length of stay distributions. This is carried out by employing a phase-type survival tree, previously used in conjunction with a standard Coxian phase-type distribution to model just a single

stage of care [15], to group patients into cohorts based on the length of stay in their initial hospital spell. Differently to the research in [15], the survival tree is constructed using a Coxian phase-type distribution with two absorbing states, to reflect the significantly different length of stay distributions for both death and discharge within each hospital stage. Once the survival tree is used to categorise elderly patients according to length of stay at the initial hospital stage, the conditional Coxian phase-type distribution is employed for each cohort separately during the subsequent stages, taking into account each individual's length of stay, at both the previous and current stages, in the determination of the transition rate parameters for the current stage. Previous research [6] has shown the conditional Coxian phase-type distribution to outperform the standard Coxian phase-type distribution when considering a number of stages within an overall aggregate system, for a population assumed to be homogeneous. This paper presents the phase-type survival tree as a front-end for this composite methodology, able to account for a heterogeneous population. The primary results of the research presented are the predictive intervals for all combinations of stage/cohort and may be found in Table 3.

The survival tree is successful in partitioning elderly patients such that those in different cohorts have a significantly different distribution for length of stay at the initial hospital stage. This is shown through the illustration of visibly disparate survival curves for each cohort. The methodology is verified through the comparison of the median from a set of simulated survival times for each fitted distribution with the 95% confidence interval for the empirical median of the represented length of stay data. Out of a total of 56 fitted distributions spanning hospital (for patients who leave through both death and community discharge) and community care stages, 53 fitted median values fall within the 95% confidence interval for the empirical median. The remaining three fitted distributions do not succeed in this respect simply due to a lack of data. The methodology has subsequently been inverted to allow predictive intervals on length of stay to be calculated for new patients entering a hospital department.

A number of assumptions made in the presentation of this research. The reasons for patient admission, whilst usually likely to have an effect on an individual's duration of stay in hospital, have been deliberately omitted from consideration. This is because the application of the research is in response to the problem surrounding the increase of elderly admissions and readmissions to hospital in a general sense, regardless as to whether each patient's subsequent readmission is for a related medical issue. In this respect, the study has been endorsed by an experienced geriatrician, highlighting the need for hospital managers to obtain a greater understanding of the movement of elderly people

between hospital and community care. A further assumption is that all patient admissions have occurred from the starting date of the study, with patients having no related hospital admissions before this point. An extension to the methodology, incorporating censoring, is to be sought to help towards alleviating this assumption. As further work, patient information may be included through the incorporation of covariates directly into the conditional Coxian phase-type distribution, thereby enabling more patient-centred predictions to be made. Additionally, this work may be extended to account for the precise type of community care that elderly patients are discharged into, for example, residential care or the patient's own home.

The current research has the potential to become a decision support tool to allow hospital managers to accurately predict when a given patient is likely to leave hospital upon entering the department, or when they are likely to be readmitted to hospital upon their discharge to the community. This would make it possible to reduce, or even eliminate, the negative effects of bed-blocking, whereby elderly patients are fit enough to be discharged from hospital, but are unable to do so due to a lack of organisation of a suitable place in community care. Additionally, alternative measures of community care may be put in place, once a given elderly patient has been discharged from hospital, in time for when they are expected to be readmitted, so that the hospital readmission may be avoided altogether. As a next step, the adaptation of the methodology presented in this research may be tested as part of a trial for a localised area, under the proviso that enough data in each of the cohorts can be obtained. With the data set in the current research taken over a period of 14 months, it may be necessary to use a wider time interval to ensure that enough observations are present for the cohorts which are determined.

The benefit of using the conditional Coxian phase-type distribution is its ability to additionally take into account the length of stay experienced by elderly patients at the previous stage of care, in the determination of transition rates for the current stage. However, the assumption that all elderly patients exhibit a similar survival distribution, and as such are homogeneous with respect to length of stay, is often not valid. The novel incorporation of the phase-type survival tree as a front-end component to the conditional Coxian phase-type distribution, results in the estimation of more accurate rate parameters, in turn allowing more precise and representative predictions to be made for a population which is heterogeneous in nature. Predictive intervals may then be calculated through the estimation of an interval based on the median of simulated times, generated from each distribution corresponding to both a stage of care and cohort. This would allow hospital managers to accurately predict when an elderly patient is likely to leave hospital, reducing the

negative effects of bed-blocking, and also to predict when an elderly patient is likely to require readmission to hospital, meaning that an alternative source of community care may be put in place, thereby avoiding readmission altogether. Implementation of the presented methodology, with a view to addressing both of the aforementioned scenarios, has the potential to lead to a more effective allocation of both hospital beds and staff, thereby meaning that vital resources may be saved.

References

1. Parry A (1996) An age-related service revisited. In: Go with the Flow: A system approach to healthcare planning, pp 127–130
2. El-Darzi E, Vasilakis C, Chaussalet TJ, Millard PH (1998) A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. Health Care Manag Sci 1(2):143–149
3. Marshall AH, McClean S (2003) Conditional phase-type distributions for modelling patient length of stay in hospital. International Transactions in Operational Research
4. Marshall AH, Vasilakis C, El-Darzi E (2005) Length of stay-based patient flow-models: recent developments and future directions. Health Care Manag Sci 8(3):213–220
5. Maguire PA, Taylor IC, Stout RW (1986) Elderly patients in acute medical wards: factors predicting length of stay in hospital. Br Med J (Clin Res Ed) 292(6530):1251–1253
6. Gordon AS, Marshall AH, Cairns KJ (2016) A conditional approach for modelling patient readmissions to hospital using a mixture of Coxian phase-type distributions incorporating Bayes' theorem. Stat Med 35(21):3810–3826
7. Garg L, McClean S, Meenan B, Millard PH (2011) Phase-type survival trees and mixed distribution survival trees for clustering patients' hospital length of stay. Informatica 22(1):57–72
8. Ross SM (2006) Stochastic processes. Wiley, New York
9. Neuts M (1981) Matrix-geometric solutions in stochastic models: an algorithmic approach. John Hopkins University Press, Baltimore
10. Marshall AH, McClean SI (2004) Using Coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. Health Care Manag Sci 7(4):285–289
11. Marshall AH, Mitchell H, Zenga M (2014) Modelling the length of stay of geriatric patients in Emilia-Romagna hospitals using coxian Phase-Type distributions with covariates. Adv Latent Variables 127–139
12. Payne K, Marshall AH, Cairns KJ (2011) Investigating the efficiency of fitting Coxian phase-type distributions to health-care data. IMA J Manag Math
13. Faddy MJ (1998) On inferring the number of phases in a Coxian phase-type distribution. Stoch Model 14(1-2):407–417
14. McClean SI, Garg L, Barton M, Fullerton K (2010) Using mixed phase-type distributions to model patient pathways. In: 23rd IEEE international symposium on computer-based medical systems, pp 172–177
15. Gordon AS, Marshall AH, Zenga M (2016) A discrete conditional phase-type model utilising a survival tree for the identification of elderly patient cohorts and their subsequent prediction of length of stay in hospital. In: 29th IEEE internations symposium on computer-based medical systems, pp 259–264
16. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control
17. Ventura C (2014) Models for predicting readmissions in heart failure patients: a comparison between Lombardia and England, PhD thesis
18. Ypma T (1995) Historical development of the Newton-Raphson method. SIAM Rev 37(4):531–551
19. Xie H, Chaussalet TJ, Millard PH (2005) A continuous time Markov model for the length of stay of elderly people in institutional long–term care. J R Stat Soc A Stat Soc 168.1:51–61