Trend Analysis of Length of Stay Data via Phase-type Models

Truc Viet Le, Nanyang Technological University, Singapore*

Chee Keong Kwoh, Nanyang Technological University, Singapore

Kheng Hock Lee, Singapore General Hospital, Singapore

Eng Soon Teo, Nanyang Technological University, Singapore

ABSTRACT

The populations in many developed countries throughout the world are aging rapidly and the number of geriatric patients is expected to rise steeply in those countries. This will exert greater pressures on the management of hospital resources as a result. Hospital length of stay (LOS) is an important indicator of hospital activity and management because of its direct relation to resource consumption. Planning of hospital resources according to identified trends of LOS is, thus, an effective way to meet such future needs. In this paper, we propose a method to analyze the temporal trends of LOS based on the Coxian phase-type distributions, a special type of continuous-time Markov process. By fitting and regressing the probabilities of discharge from each phase of the distribution on time, we have found a growing trend in the proportion of long-staying patients in our sample of stroke patients from a general hospital in Singapore. We compare the yearly, quarterly and monthly trends over the same period to see the common pattern. The datasets were also robustified by bootstrapping to aid the analysis.

Keywords: length of stay, hospital planning, trend analysis, phase-type, Markov processes, healthcare, information systems.

INTRODUCTION

Hospital length of stay (LOS) is often used as a reliable proxy for measuring the consumption of hospital resources because it exhibits a strong correlation with resource consumptions (Librero*, et al.*, 2004). LOS is also an easily available indicator of hospital activity and is used for various purposes such as management of hospital care, quality control, and hospital planning. Therefore, LOS is a key performance indicator for hospital management and a key measure of efficiency of the healthcare system currently implemented (Kulinskaya*, et al.*, 2005). According to Lee *et al.* (2005), “comprehensive and accurate information about inpatient LOS should be a high priority for health planners and administrators in the strategic planning and deployment of financial, human and physical resources.”

The mean of LOS is often used as a benchmark for the consumption of resources because of its readiness and good relation with raised costs. However, the common approach of averaging LOS is only a localization measure of the variable and it would be misleading if the underlying distribution is not symmetrical (Vasilakis & Marshall, 2005). The empirical distribution of LOS is established to be positively skewed with a heavy right tail. It is also well-known to contain many outliers and significantly vary between homogeneous groups of patients. This heterogeneity of LOS has posed serious problems for statistical analysis and has limited the use of inference techniques based on normality assumptions.

Previously, optimizing the use of LOS as an indicator of care has been attempted through various mathematical models. LOS data are conventionally analyzed using the non-parametric methods of survival analysis (Li, 1999). Survival analysis typically uses LOS data as a vehicle to study the effects of patient features on survival time (which is equivalent to length of stay in hospital), such as the differential effects of social attributes (Kwoh*, et al.*, 2009), or discharge destinations on LOS (Xie*, et al.*, 2006). Mixture models have been used by Atienza *et al.* (2008) to approximate the distributions of LOS of certain Diagnosis-Related Groups (DRG) . Abbi *et al.* (2008) further used mixture distributions to categorize patients into homogeneous groups . Compartmental models have also been applied to model LOS as flow of patients through various wards (Vasilakis & Marshall, 2005). Recently, the Coxian phase-type distributions, a special type of continuous-time Markov process, have been shown to be able to model LOS distributions accurately with many useful applications such as identifying patient groups and estimating associated costs (Faddy & McClean, 1999; A. Marshall & McClean, 2004; A. H. Marshall*, et al.*, 2007). See Fackrell (2009) for a thorough literature review of phase-type distributions and their applications to modeling LOS.

However, practical matters such as analyzing temporal tends of LOS have not yet been addressed using any of the above-mentioned methods. This paper proposes a novel method to analyze trends of LOS using the Coxian phase-type distributions and simple regression analyses. The distribution has the ability to stochastically model the transitions of patients through different progressive stages of stay. This allows for the comparison of the discharge probabilities from each stage over the periods of interest. A trend can be identified by inspecting any patterns of change in these probabilities over time. This can be done by regressing these probabilities on time as the sole explanatory variable.

MATERIALS AND METHODS

Dataset

The analyses were performed on lengths of stay of 4,097 inpatients discharged from the Singapore General Hospital (SGH) in the duration of four years, from 2004 to 2007. In addition, all these patients were diagnosed with AN-DRG (Australian National Diagnosis-Related Group V3.2 (1997)) 37 – “Cerebrovascular Disorders except TIA with CC” – a major stroke DRG. DRG is a classification system of episodes of hospitalization with clinically recognized definitions, where it is expected that patients in the same group consume similar quantities of resources as a result of a similar care process. When DRG’s are built, length of stay is often used as the outcome variable because it exhibits a strong correlation with resource consumptions (Librero, et al., 2004).

The reason for choosing a sample from a specific DRG is to avoid the large variance in LOS that would be associated with heterogeneous groups of DRG. LOS within a DRG group is assumed to be homogeneous and specific to that group. Moreover, stroke is a disease of special interest to us because it characteristically leads to a diverse hospital population in terms of clinical management. The neurological deficit occurring after stroke may be a transient phenomenon or may cause various degrees of disability. Therefore, these observations pose interesting challenges in the modeling of stroke-related LOS data. Also, as the prevalence of the disease is relatively high (stroke is the third most common cause of death in most developed countries according to Gubitz (2000)), it will be of much use to study the lengths of stay and resource consumption of stroke-related patients.

Table 1 gives a descriptive statistics of the LOS of the group of stroke-related patients. The distribution of LOS observed from the table is highly right-skewed (given by a high positive skewness) with a very high peak at the mode (given by a high kurtosis) which is typical of LOS distributions. The asymmetric and skewed nature of LOS distribution is further illustrated in the LOS histogram (black solid line) shown in Fig. 2.

Table1. Descriptive statistics of LOS of years 2004–2007

n	4,097
Mean	11.530
Std. Deviation	11.958

Skewness	1.982
Kurtosis	5.216
Min	0
1st Quartile	3
Median	7
3rd Quartile	16
Max	97

Coxian Phase-type model

Phase-type (PH) distributions, introduced by Neuts (Miller, 1983), describe the time spent in the transient states of a finite-state continuous-time Markov chain with one absorbing state until absorption. A state is *transient* if once it has been reached, the probability of returning to it is less than one, and a state is *absorbing* if once it has been reached, it is impossible to leave the state, i.e., the process stops. In PH distributions, there is a single absorbing state and the stochastic process starts in a transient state.

Distributions of this form have significant generality; they include the exponential distributions, Erlang distributions, and mixture of exponential distributions (Breuer & Baum, 2006). In fact, they can approximate any non-negative distribution arbitrarily closely with sufficiently large number of states (Fackrell, 2009).

Let $Z \coloneqq \inf\{t \geq 0 : X_t = n + 1\}$ be the random variable of the time until absorption in state $n + 1$. The distribution of Z is called phase-type distribution with *representation* $(\alpha, \mathbf{T})$. The dimension n is called the *order* of the distribution and the states $\{1, 2, \cdots, n\}$ are called *phases*.

Coxian phase-type is a special type of phase-type distributions (Cox, 1955). Coxian PH has the transient states *ordered* with the process starting in the first state and then either progresses through the phases sequentially or enters into the absorbing state. For an n-phase Coxian PH model, there are $2n - 1$ parameters to estimate. The process is illustrated in Fig. 1. Note that in the figure, the absorbing state is denoted as state “0" and the entering arrow with probability 1.0 means that the process always starts at the first phase.

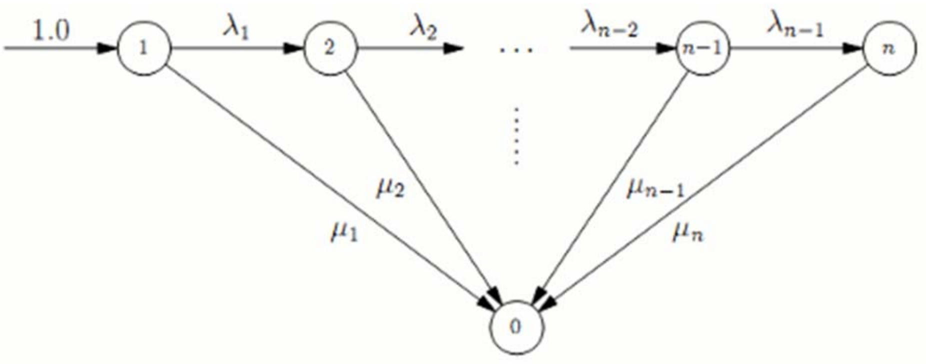

Figure 1. Illustration of a general Coxian phase-type distribution

The Coxian phase-type distributions describe the probability that the process is still active at time t. Let $\{X(t) \mid t \geq 0\}$ be a Markov process in continuous time over the states $\{1, 2, \cdots, n, n+1\}$ with $\{X(0) = 1\}$, and infinitesimal transition probabilities, for $i = 1, 2, \cdots, n-1$, are given by

$$\Pr\{X(t+\delta t) = i+1 \mid X(t) = i\} = \lambda_i \delta t + o(\delta t), \tag{1}$$

and for $i = 1, 2, \cdots, n$:

$$\Pr\{X(t+\delta t) = n+1 \mid X(t) = i\} = \mu_i \delta t + o(\delta t). \tag{2}$$

The states $1, 2, \cdots, n$ are *transient* and state $n+1$ is an *absorbing* state. The parameters λ_i's describe the transitions through the ordered transient states $\{1, 2, \cdots, n\}$ and the parameters μ_i's describe the transitions into the absorbing state, which is depicted as state "0" in Fig. 1. Hence, over a sufficiently small interval of time δt, the probability of a transition is approximately proportional to the interval's duration. The distributions can be represented in matrix notation where the probability density function is given by

$$f(t) = \alpha \exp(\mathbf{T}t)\, \eta = \alpha \sum_{k=0}^{\infty} \frac{\mathbf{T}^k t^k}{k!} \eta, \tag{3}$$

where

$$\alpha = (1\ 0\ 0\ \cdots\ 0\ 0), \tag{4}$$

$$\eta = -\mathbf{T} \cdot \mathbf{1} = (\mu_1\ \mu_2\ \cdots\ \mu_n)^{\mathrm{T}}, \tag{5}$$

and $\mathbf{T}$ is the matrix of transition rates between the phases and is given by

$$\mathbf{T} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\lambda_{n-1}+\mu_{n-1}) & \lambda_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & -\mu_n \end{pmatrix}. \tag{6}$$

See Breuer & Baum (2006) for more detailed description.

In terms of length of stay in hospital, the patients are admitted into phase 1 and their movement through the transient phases could be thought of as their progression through the stages of treatment and care in hospital, with the absorbing phase representing discharge or death. Absorption from the earlier phases would represent the end of treatment and care for a patient with a relatively short duration of stay, while absorption from later phases would represent the end of stay for those needing more treatment and care.

Fitting the model

The parameters of the Coxian PH distribution, μ_i's and λ_i's, can be estimated from the LOS data of the group of stroke-related patients. R program (R Team, 2009) was used to implement an optimization routine based on the expectation-maximization (EM) algorithm by Asmussen *et al.* (1996) to estimate the parameters. The fitting was done in a trial-and-error fashion in which the simplest model of two phases was fitted first, the process continues by adding one more phase each time until a certain degree of fitness was obtained. We start with a two-phase Coxian PH model because a one-phase model corresponds to an exponential distribution and, hence, not considered (because the distribution of LOS clearly has a non-zero mode.) A sufficient model is one that is capable of approximating the original distribution to a certain level of fidelity but at the same time is not too complex, i.e., it should not *overfit* the data. For this reason, the main criterion chosen for model selection is the Akaike Information Criterion (AIC) (Akaike, 1974). Besides AIC, we also make use of other auxiliary goodness-of-fit scores for the purpose of comparison, such as the log-likelihood (logLik), R^2 and Kullback–Leibler divergence (KL) (Kullback & Leibler, 1951). Hence, a preferred model is one that maximizes the log-likelihood and R^2 and at the same time minimizes AIC and KL.

Table 2 presents the results of fitting the data to the Coxian PH model, starting from a two-phase model and ending at a six-phase one. The phases are shown in rows and the goodness-of-fit scores are shown in columns. From the table, it is concluded that a five-phase model is the most appropriate. This is because in the six-phase model, AIC begins to increase. Besides AIC, other scores also support this conclusion, R^2 and KL don't seem to improve as more phases are added.

Table 2. Goodness of fit of the Coxian PH model for LOS of years 2004–2007

	logLik	$R^2(\%)$	AIC	KL
2	-13217.810	99.99993	26441.620	0.146
3	-13007.382	99.99997	26024.764	0.061

4	-12987.552	99.99997	25989.104	0.056
5	-12984.271	99.99997	25986.541	0.055
6	-12982.997	99.99997	25987.993	0.055

Table 3 summarizes all the estimated parameters of the Coxian PH models (from two to six-phase) for the purpose of comparison. The parameters are shown in rows and the phases are shown in columns. An “NA” value in the table indicates that the parameter is not applicable to a particular model and a “0” value means that the estimated parameter has a value of zero.

Table 3. Parameters estimates of the Coxian PH distributions for LOS of years 2004–2007

	2	3	4	5	6
λ_1	0.052	0.538	0.438	0.420	0.414
λ_2	NA	0.340	0.204	0.183	0.176
λ_3	NA	NA	0.089	0.084	0.083
λ_4	NA	NA	NA	0.308	0.387
λ_5	NA	NA	NA	NA	0.350
μ_1	0.082	0.002	0.002	0.002	0.002
μ_2	0.075	0.200	0.235	0.239	0.240
μ_3	NA	0.070	0	0	0
μ_4	NA	NA	0.159	0	0
μ_5	NA	NA	NA	0.308	0.055
μ_6	NA	NA	NA	NA	0.544

Finally, Fig. 2 shows graphically the goodness of fit of the five-phase Coxian PH model to the data. It shows the fitted curve (in grey dashed line) superimposed over the empirical histogram (in black solid line). As observed from the figure, the estimated line closely fits the histogram, thus supporting our phase-type model.

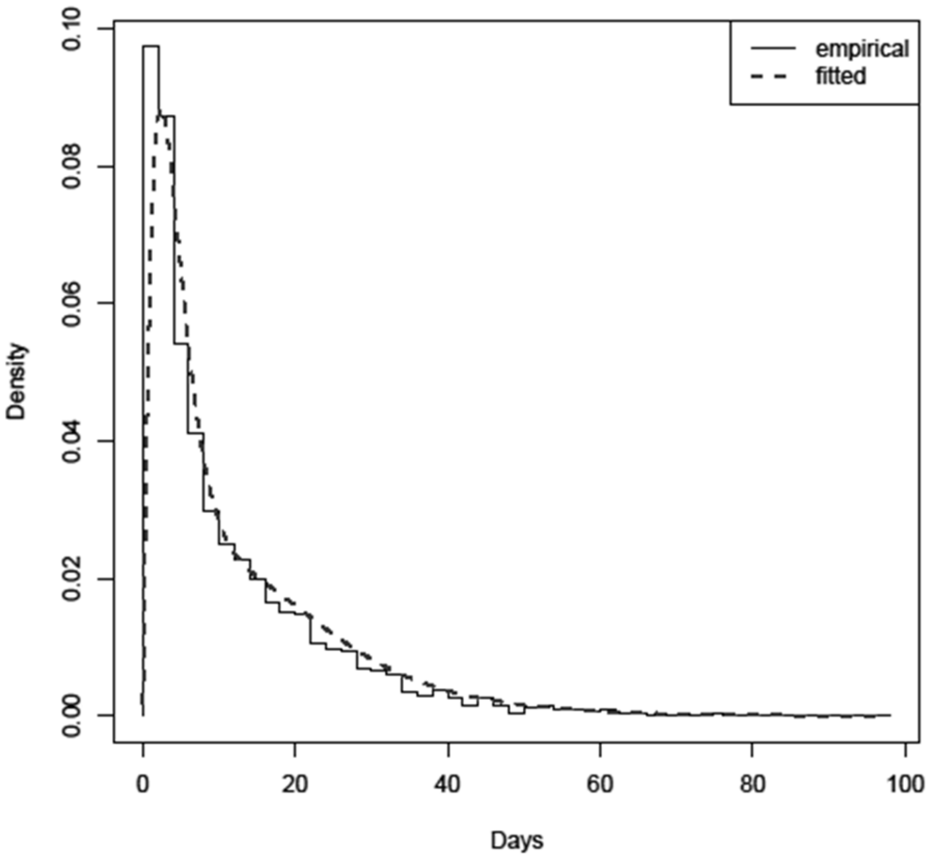

Figure 2. Illustration of the 5-phase model's goodness of fit for LOS of years 2004–2007

Analyzing the trends of LOS

To analyze the trend of LOS over the years 2004–2007, we split the LOS data into periods, fit each period's data to the Coxian PH model and compare the variations in model parameters over the periods. Specifically, we split the data into years (4), quarters (16) and months (48) and then analyze the trend of LOS over years, quarters, and months. However, a question that naturally arises is how many phases of the PH distribution should be used to fit the data in order to have a consistent comparison. We use a heuristic approach as follows to determine the most appropriate global number of phases.

We first fit the yearly LOS data and use the selection criteria as described in the previous section to determine the number of phases for each year. Table 4 shows the descriptive statistics of the yearly data. Table 5 shows the resultant goodness-of-fit scores of the model for each year. The number of phases appears in parentheses next to the year in the first column of the table.

Table 4. Descriptive statistics of LOS for each year in 2004–2007

	2004	2005	2006	2007
n	962	1,032	1,009	1,094
Mean	10.447	11.769	11.591	12.217
Std. Deviation	10.412	12.257	11.860	12.942
Skewness	2.113	1.851	1.706	2.123
Kurtosis	6.013	3.986	3.060	6.562
Min	0	0	0	0
1st Quartile	3	3	3	3
Median	7	7	7	7
3rd Quartile	13	17	17	17
Max	85	83	75	97

Table 5. Goodness of fit of the Coxian PH model for each year in 2004–2007

	logLik	R^2(%)	AIC	KL
2004(4)	-2955.487	99.99995	5924.974	0.084
2005(4)	-3272.358	99.99996	6558.716	0.085
2006(4)	-3187.754	99.99996	6389.507	0.090
2007(6)	-3538.656	99.99997	7099.311	0.104

We observe from Table 5 that in 3 out of 4 years under study, the number of phases that best fits the data is 4 phases. Therefore, we decide that 4 is the number of phases that is most appropriate for modeling and comparing the periodic data. Moreover, choosing 4 as the global number of phases has an advantage over 5 because it means that the majority of the data is not overfitted while some small degree of model accuracy could be sacrificed. In statistical modeling, the former is often more important than the latter.

The probability π_i that a patient is discharged from phase i can be calculated as follows. Let π_i be the probability that an individual departs from the system (i.e., is absorbed) from phase i. This probability can be calculated by taking the probability density formula for each phase. For example, for phase 1,

$$\pi_1 = \int_0^\infty \mu_1 e^{-(\lambda_1 + \mu_1)t}\, dt = \frac{\mu_1}{\lambda_1 + \mu_1}. \tag{7}$$

Similarly, for phase 2,

$$\pi_2 = \int_0^\infty \mu_2 e^{-(\lambda_2 + \mu_2)t} dt \int_0^\infty \lambda_1 e^{-(\lambda_1 + \mu_1)t} dt \tag{8}$$

$$= \frac{\lambda_1}{\lambda_1 + \mu_1} \cdot \frac{\mu_2}{\lambda_2 + \mu_2}.$$

In general, for an n-phase model and $1 < i < n, \pi_i$ is given by

$$\pi_i = \prod_{j=1}^{i-1} \frac{\lambda_j}{\lambda_j + \mu_j} \cdot \frac{\mu_i}{\lambda_i + \mu_i}, \tag{9}$$

and for $i = n$,

$$\pi_n = 1 - \sum_{j=1}^{n-1} \pi_j. \tag{10}$$

In the subsequent comparisons of the fitted model for each period, we define the proportion of *short-staying* patients as those who were discharged with probability $\pi_1 + \pi_2$ and the proportion of *long-staying* patients is the rest, i.e., those who were discharged with probability $\pi_3 + \pi_4$.

Bootstrapping the data

Since dividing the four years' data into periods (quarters, months) could significantly reduce the number of data points per fit and hence renders the estimates less significant. We use of bootstrapping to robustify the parameter estimates. Note that we do not bootstrap the yearly data because it is sufficiently large as shown in Table 4.

Bootstrapping is a simple but powerful random resampling with replacement method to assess the statistical estimates from the sample statistics. Being a robust validation and pruning technique, bootstrapping can be used to validate models and reduce estimation errors. According to by Efron & Tibshirani (1997), although the traditional cross-validation method is nearly unbiased, it can be highly variable. The estimation error of bootstrapping has substantially outperformed the cross-validation technique in many experiments.

In our experiment, bootstrapping is carried out as follows:

1. Given the LOS data of a quarter or month according to admission dates,
2. Generate bootstrapped LOS samples by randomly resampling with replacement from the given data. This is done m times, depending on the amount of time and computing power available.

3. Fit each bootstrapped sample to a Coxian PH distribution using the EM algorithm to obtain the model parameters.
4. Average out the parameters to obtain the final model parameters of the quarter.

Bootstrapping assumes that each data point is independently and identically distributed (i.d.d.). This can be safely assumed in our datasets. As the number of bootstrap samples, m, tends to infinity, the distributions of the fitted parameters tend to normal distributions. Without loss of generality, m is chosen to be 30 here for practical reasons. Hence, the computed parameters used in later analyses will be the *averaged* parameter of all the bootstrapped samples.

Finally, Fig. 3 summarizes the whole process of the trend analysis that has been described in this section.

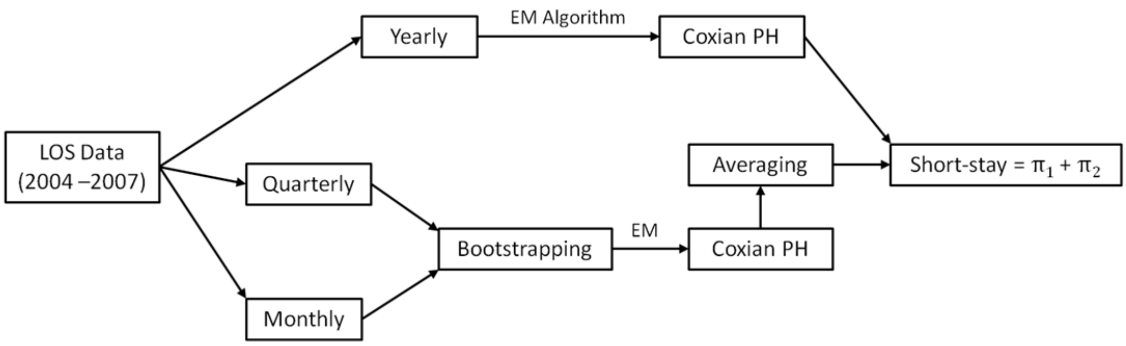

Figure 3. The flowchart describing the whole process of the trend analysis

RESULTS

The yearly trend

Table 6 summarizes the yearly probability of discharge from each phase of the Coxian PH distribution. A “0” value means that the probability of being discharged from that phase is zero.

Table 6. Changes in proportions of LOS over the years in 2004–2007

	π_1	π_2	π_3	π_4
2004	0	0.678	0	0.322
2005	0.004	0.538	0	0.458
2006	0.004	0.523	0	0.473
2007	0.007	0.484	0.350	0.159

Fig. 4 shows the plot of the proportions of short-staying patients over the years. The plot shows an emerging pattern of change: from 2004 to 2007, the proportions of patients who stay for short periods of time (i.e., those discharged from phase 1 and 2) are decreasing. In other words, the proportions of patients who stay for longer periods of time (i.e., those discharged from phase 4 and 5) are increasing.

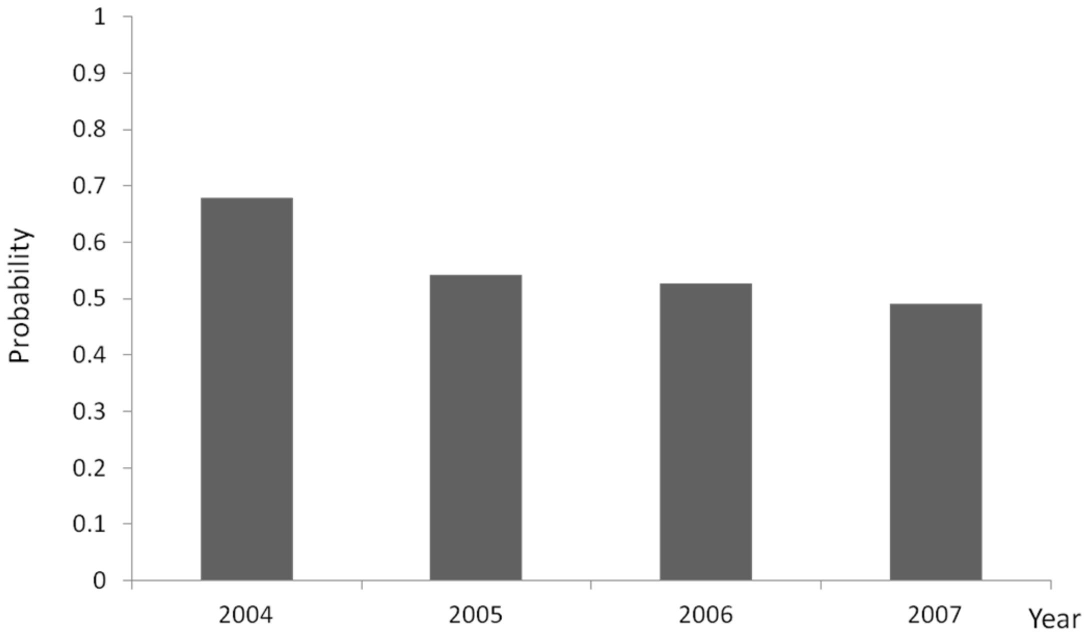

Figure 4. Graphical comparison of proportions of short-staying patients over the years 2004–2007

However, due to the small sample size in this "trend" ($n = 4$ in this case), this is not a rigorous argument. We only use it to provide a visual clue to what will follow in the next sections where we analyze the quarterly and monthly trends of LOS rigorously.

The quarterly trend

Table 7 summarizes the quarterly probability of discharge from each phase of the Coxian PH distribution of the bootstrapped data.

Table 7. Changes in proportions of LOS over the quarters in 2004–2007

	π_1	π_2	π_3	π_4
Q1	0	0.600	0.025	0.374
Q2	0	0.601	0.006	0.393
Q3	0	0.721	0	0.279
Q4	0	0.568	0.075	0.358
Q5	0	0.592	0	0.408

Q6	0.119	0.370	0.099	0.411
Q7	0	0.580	0	0.420
Q8	0	0.453	0.006	0.541
Q9	0	0.563	0.005	0.432
Q10	0.053	0.471	0.050	0.426
Q11	0	0.513	0.011	0.476
Q12	0.069	0.418	0	0.513
Q13	0.037	0.436	0	0.527
Q14	0.090	0.420	0.052	0.438
Q15	0.032	0.416	0.003	0.550
Q16	0.130	0.363	0.088	0.419

Fig. 5 shows the plot of the discharge probabilities over the 16 quarters of the bootstrapped data. The area under the curve shows the probability that a discharge is of a short-staying patient and the area above the curve shows the probability that the discharge is of a long-staying patient. The sum of the two areas must be 1 at all times. In addition, a linear trendline is also plotted in Fig. 5 (to be elaborated later).

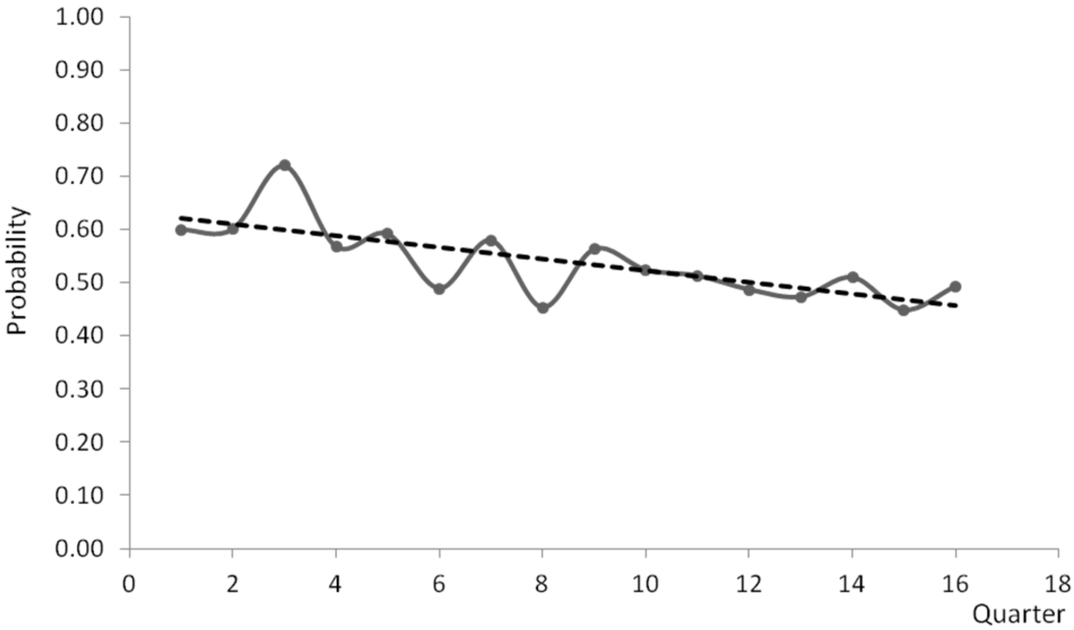

Figure 5. The probability of discharge over the 16 quarters of 2004–2007 (the dashed line shows a linear trendline)

Fig. 5 shows an apparent declining trend in the probability of discharge of short-staying patients over the 16 quarters. This confirms what was observed in the yearly trend in Fig. 4. We analyze this trend by performing a linear regression analysis $y = \beta_0 + x\beta + \epsilon$ on the trend by the least squares method. The dependent variable, y, is the probability of discharge of short-staying patients, which is represented by the solid curve in the plot of Fig. 5. The sole predictor, x, is time (i.e., the 16 quarters). β_0 is the intercept, β is the regression coefficient and ϵ is the residual term, which is assumed to be normally distributed with mean 0 and standard deviation 1. Table 8 shows the results of the regression analysis.

Table 8. Results of regression analysis on the quarterly trend ($R^2 = 53.82\%, n = 16$)

	Coefficients	Std. error	t-stat	p-value
β_0	0.6315	0.0262	24.1001	8.47×10^{-13}
β	-0.0109	0.0027	-4.0393	0.0012

Table 8 shows that both the coefficient estimates β_0 and β are significant at 95% confidence interval (both p-values are less than 5%). A negative β coefficient confirms a declining trend of short-staying patients. Besides, the R^2 coefficient is also relatively significant at 53.82%. The trendline plotted in Fig. 5 is actually calculated using these coefficient estimates.

Here we chose simple linear regression to analyze the trend for the reason of model parsimony. However simple, the linear regression model was able to explain for more than 50% of the variability in the data, which is quite sufficient for our purpose. Nevertheless, in practice, more complex models could be used to find out a trend if the probability curve exhibits highly non-linear characteristics.

The monthly trend

We skip presenting the table summarizing the discharge probabilities in this section to save space (because, otherwise, we would need a table with at least 48 columns to do that.) Fig. 6 shows the plot of the discharge probabilities over the 48 months of the bootstrapped data. It, again, shows a declining trend in the probability of discharge of short-staying patients. However, the trend here exhibits some degree of volatility that is possibility due to seasonal fluctuations. A linear trendline similar to that of Fig. 5 is also plotted. The trendline has a downward slope as expected even though it is less steep than the one in Fig. 5.

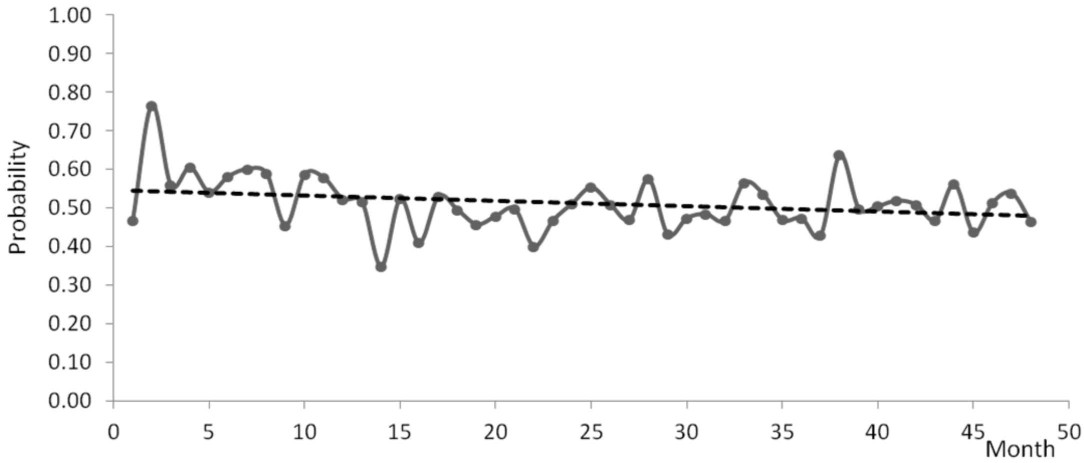

Figure 6. The probability of discharge over the 48 months of 2004–2007 (the dashed line shows a linear trendline)

Similar to the quarterly trend analysis, we analyze the monthly trend by a simple regression model and the results are presented in Table 9. From the table, it is observed that the coefficient estimate of β_0 is highly significant while the estimate of β is relatively insignificant, even though it is close to the 5% threshold (p-value of β is 5.3%). The R^2 goodness-of-fit indicator is also quite weak at 7.9% as compared to the quarterly results. However, we argue that this is due to seasonal fluctuations that are often observed in monthly data (as can be seen in Fig. 6) and the claim of the trend still holds, which is indicated by the negative value of β coefficient.

Table 9. Results of regression analysis on the monthly trend ($R^2 = 7.9\%, n = 48$)

	Coefficients	Std. error	t-stat	p-value
β_0	0.5467	0.0198	27.6015	2.85×10^{-30}
β	-0.0014	0.0007	-1.9863	0.0530

CONCLUSION

It has been predicted by the U.S. Census Bureau (2001) that the percentage of the elderly comprising the total population will more than double over the next 50 years. This will certainly increase the number of hospital inpatients, especially those with stroke-related diseases which are prevalent among them. As a result, the burden levied on the healthcare sector in terms of managing and planning of existing and future resources will become heavier. Hence, efficient resource allocations based on temporal trends of LOS will become a more appropriate method. LOS is a reliable indicator of hospital resource consumption and has been traditionally used for the purpose of rationing resources and facilities. However, due to its extreme asymmetrical nature, the conventional use of simple statistic such as average length of stay is misleading and inaccurate. The dependence on this indicator by hospital management and policy making would result in inefficiency.

The Singapore General Hospital (SGH) is the largest tertiary hospital in Singapore and is in many ways typical of a tertiary general hospital in a developed country. This paper illustrates the use of the Coxian phase-type (PH) distributions in modeling and analyzing trends of LOS of stroke-related patients from such a hospital. The Coxian PH model has been shown to be able to capture satisfactorily the variance of LOS. The clinical care of stroke patients occurs in phases, such phases are distinctive in a patient's journey in the hospital from the emergency room to the eventual discharge. The phases of the Coxian PH model therefore provide more information than the average LOS over time. The phases that appear in a Coxian PH model may reflect the transitional phases of patient care from acute management to convalescence. It may also be affected by the availability of resources for step-down care.

Our analysis shows that there has been a pattern of change over recent years in LOS. This change is captured by the change in the parameters of the Coxian PH model over the periods. This gives a novel perspective to the study of trends of LOS: comparing the probability of discharge from each phase through the periods. The paper has shown that, for our particular dataset, the proportions of patients discharged after short durations of stay have been declining, or, in other words, the proportions of those discharged after longer durations have been rising. This supports the widely-held perception that recent efforts in aiming at reducing LOS through early and aggressive interventions had been negated by the lack of step-down care facilities in the community (Yacob, 2008). In such a case, the bottle-neck of the care process lies at the final phases of hospitalization. Therefore, putting efforts and resources to resolve problems at this phase of care will probably be the more cost-effective way of improving the overall system efficiency.

While it would require more data to validate the identified trend, we believe that healthcare planners can benefit from the proposed method of studying trends of LOS. The improved use of indicators such as LOS will result in better understanding and anticipation of changing healthcare needs. This in turn will enable more efficient allocations of healthcare resources.

ACKNOWLEDGEMENT

This work was supported by the Biomedical Engineering Research Centre seed funding (SPF 022-2005) of the Nanyang Technological University (NTU). The study was approved by the Institutional Review Board (IRB #43/2008) of the Singapore General Hospital (SGH). We would like to thank Ms. Chee Thong Gan, Deputy Director of Allied Health Division, SGH, and the IT department of SGH for their help in this study.

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. *Computer-Based Medical Systems, IEEE Symposium on, 0*, 524-529.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *{IEEE} Transactions on Automatic Control, 19*(6), 716-723.

Asmussen, S. o. r., Nerman, O., & Olsson, M. (1996). Fitting Phase-Type Distributions via the EM Algorithm. *Scandinavian Journal of Statistics, 23*(4).

Atienza, N., García-Heras, J., Muñoz-Pichardo, J. M., & Villa, R. (2008). An application of mixture distributions in modelization of length of hospital stay. *Statistics in Medicine, 27*(9), 1403-1420. doi: 10.1002/sim.3029

Breuer, L., & Baum, D. (2006). *An Introduction to Queueing Theory and Matrix-Analytic Methods*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Bureau, U. S. C. (2001). An Aging World: International Population Reports: U.S. Census Bureau.

Cox, D. R. (1955). A use of complex probabilities in the theory of stochastic processes. *Mathematical Proceedings of the Cambridge Philosophical Society, 51*(02), 313-319. doi: doi:10.1017/S0305004100030231

Efron, B., & Tibshirani, R. (1997). Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association, 92*(438), 548-560.

Fackrell, M. (2009). Modelling healthcare systems with phase-type distributions. *Health Care Management Science, 12*(1), 11-26. doi: 10.1007/s10729-008-9070-y

Faddy, M. J., & McClean, S. I. (1999). Analysing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry, 15*(4), 311--317.

Gubitz, G. (2000). Extracts from "Clinical Evidence": Acute ischaemic stroke. *BMJ, 320*(7236), 692--696.

Kulinskaya, E., Kornbrot, D., & Gao, H. (2005). Length of stay as a performance indicator: robust statistical methodology. *IMA Journal of Management Mathematics, 16*(4), 369-381.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 49--86.

Kwoh, C.-K., Lee, K.-H., & Le, T.-V. (2009, dec.). *Using survival models to analyze the effects of social attributes on length of stay of stroke patients.* Paper presented at the Biomedical and Pharmaceutical Engineering, 2009. ICBPE '09. International Conference on.

Lee, A. H., Gracey, M., Wang, K., & Yau, K. K. W. (2005). A robustified modeling approach to analyze pediatric length of stay. *Annals of Epidemiology, 15*(9), 673--677.

Li, J. (1999). An application of lifetime models in estimation of expected length of stay of patients in hospital with complexity and age adjustment. *Statistics in Medicine, 18*(23), 3337--3344.

Librero, J., Marin, M., Peiro, S., & Munujos, A. V. (2004). Exploring the impact of complications on length of stay in major surgery diagnosis-related groups. *International Journal for Quality in Health Care, 16*(1), 51-57.

Marshall, A., & McClean, S. (2004). Using Coxian Phase-Type Distributions to Identify Patient Characteristics for Duration of Stay in Hospital. *Health Care Management Science, 7*, 285-289.

Marshall, A. H., Shaw, B., & McClean, S. I. (2007). Estimating the costs for a group of geriatric patients using the Coxian phase-type distribution. *Statistics in Medicine, 26*(13), 2716--2729.

Miller, D. R. (1983). Matrix-geometric solutions in stochastic models: An algorithmic approach, by Marcel R. Neuts. *13*(2), 311--312.

R Team, D. C. (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Vasilakis, C., & Marshall, A. H. (2005). Modelling Nationwide Hospital Length of Stay: Opening the Black Box. *The Journal of the Operational Research Society, 56*(7), pp. 862-869.

Xie, H., Chaussalet, T. J., & Millard, P. H. (2006). A model-based approach to the analysis of patterns of length of stay in institutional long-term care. *IEEE Trans Inf Technol Biomed, 10*(3), 512-518.

Yacob, H. (2008). Healthcare policy and rising medical cost. *Singapore Parliament Hansard, 84, Parliament no. 11, session 1, sitting no. 9*.