21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

Determining the Optimal Number of Beds in the Subacute Section of a Large Hospital

R. García-Flores [a], **Ross Sparks** [b], **Debbie Munro** [c] **and Alan McCubbin** [c]

[a]*CSIRO Digital Productivity Flagship, Private Bag 33, Clayton South, VIC, Australia, 3169*
[b]*CSIRO Digital Productivity Flagship, PO Box 52, North Ryde, NSW, Australia, 1670*
[c]*Austin Health, PO Box 5555, Heidelberg, VIC, Australia, 3084*
Email: Rodolfo.Garcia-Flores@csiro.au

Abstract: Austin Health (AH) currently operates inpatient Continuing Care services across two satellite campuses, and faces the problem of calculating the optimal number of beds needed in its subacute wards. It is inefficient if a patient needs to wait in an acute ward for a bed in a subacute ward, just as is operating too many beds in the subacute ward to always meet demand. In recent years the demand for acute beds has increased, creating pressure for faster patient movements and more admissions at times of high demand, with the consequent need to determine best-performing bed configurations. Additional constraints, related mostly to availability of medical resources, were of concern to AH staff and were considered when developing the model. The subacute bed allocation problem is significant because it cannot be formulated in closed form using simple probability distributions, but demands the use of actual variable data on admissions and separations to ensure a reliable result. The solution approach we used to tackle the problem is based on the combined use of the cross-entropy method for optimisation. It uses the simulation of subacute ward occupation and demand using a parametric bootstrap to generate data to solve this problem. We used a simulation model to represent the six wards under study and the dynamic relationships that describe this system. To obtain the optimal bed configurations, we use the cross-entropy method for optimisation. This is a modern optimisation method whose working principle is based on the fact that cross-entropy divergence can be used as a measure of closeness between two sampling distributions. Optimisation by cross-entropy estimates a sequence of parametric sampling distributions that converges to a distribution with probability mass concentrated in a region of near-optimal solutions. We used a parametric bootstrapping approach to generate the admission data that is used as an input to the optimiser. The expected result is a slight increase in the number of existing beds. We justify the effectiveness of the proposed approach for determining the optimal number of beds on the grounds that the actual results and the general behaviour of the optimisation software in its current version match the intuition of hospital staff on the behaviour of the system.

Keywords: *e-Health, subacute bed allocation, cross-entropy optimisation, parametric bootstrap, hospital capacity planning*

1. INTRODUCTION

In Australia, the number of people aged 65 and over is projected to increase to more than double the current 2015 figures by 2054-2055 [The Treasury, Australian Government, 2015]. As Australians will live longer and continue to have one of the longest life expectancies in the world, there will be fewer people of traditional working age compared with the very young and the elderly. This trend is already visible, with the number of people aged between 15 and 64 for every person aged 65 and over having fallen from 7.3 people in 1974-75 to an estimated 4.5 people today. By 2054-55, this is projected to nearly halve again to 2.7 people.

The allocation of limited and costly health services to an ageing population is a tremendous challenge recognised by government and industry alike: see, for example, Rosen et al. [2010] and Motorola [2010]. The need for increased data collection, smarter decision making and improved logistic processes is widespread and urgent, as current policies and practices clearly lead to increasing costs, less effective healthcare, and may even put patient safety at risk. In this paper, we contribute to ameliorate this situation by tackling the problem of calculating the optimal number of beds needed in the subacute wards of a major hospital.

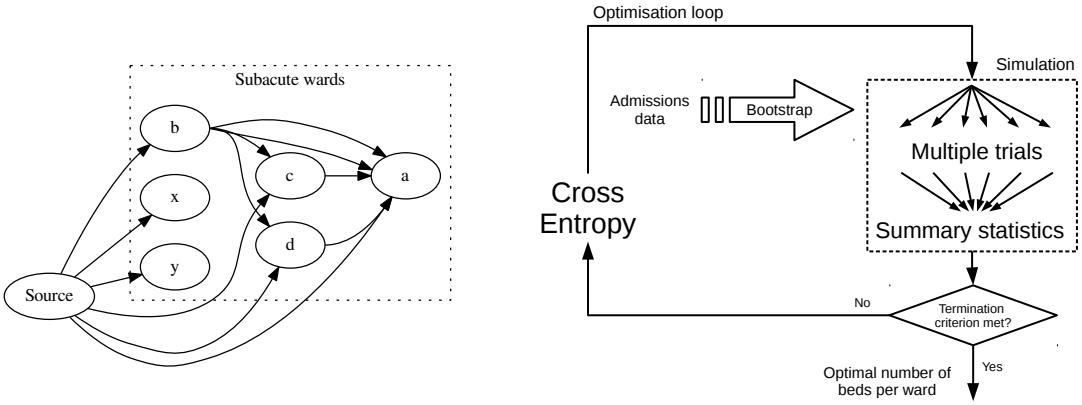
Austin Health (AH) currently operates inpatient Continuing Care services across two satellite campuses that are separate from the acute hospital: the Heidelberg Repatriation Hospital (HRH), which provides Geriatric Evaluation and Management and Rehabilitation Care types, and the Royal Talbot Rehabilitation Centre (RTRC), which provides Rehabilitation Care only. Apart from the opening of a 24 bed geriatric ward at the HRH in 2008 and the closure of five rehabilitation beds on the RTRC in 2013, the number and configuration of beds has remained largely unchanged. In recent years the demand for acute beds has increased, creating pressure for faster patient movements and more admissions at times of high demand. Additional constraints, some of which are specific to each campus, were of concern to AH staff. Most of these relate to access to medical resources, as for example support services such as pathology, radiology or staff's expertise, but others were related to funding, transportation or general bed access and availability. Although not all of these were considered in the development of the optimisation model, the possibility of doing so was discussed with AH staff.

Different approaches have been used in the past to solve this problem. Bachouch et al. [2012] use an integer linear model for hospital bed planning for acute and elective patients. This study emphasised the role of restrictions such as patient assignment to double rooms with consideration to compatibilities between pathologies, or having patients with contagious diseases in separate rooms. Some of the assumptions of this study are not easy to justify, e.g., the length of stay is known, and that there are no changes of bed through the hospitalisation. Simulation has also been commonly used to tackle this problem. Wang et al. [2011] developed a queuing model for emergency and acute care bed allocation for various hospitals to answer the question of how to distribute beds across hospital wards in order to maximise the quality of patient care. Zhang et al. [2012] present a combined simulation and optimisation approach to determine long-term bed capacity levels in a Canadian hospital over a multi-year planning horizon, and compare it to a fixed-ratio approach used in practice. The data is assumed to follow a Poisson distribution for arrivals and a Weibull distribution for length of stay (LOS), and a bisection method is selected as the optimisation algorithm. Other papers adopt a queue-theoretical approach. In de Bruin et al. [2010], an Erlang loss model was introduced to test the validity of the current guideline of 85% occupancy in Dutch hospitals. In contrast to de Bruin et al. [2010], our study considers the difference in arrival patterns by day of the week, including weekends.

In this paper, we present a solution approach based on the combined use of the cross-entropy (CE) method for optimisation [Rubinstein, 1997, 1999; Rubinstein et al., 2013], the simulation of subacute ward occupation, and parametric bootstrapping of the data. To the best of our knowledge, CE has only been used in the health sector to determine thresholds on histograms to analyse staining in epithelial cells in Neves et al. [2014]. We used a simulation model to represent the six wards under study and the dynamic relationships that describe this system. To obtain optimal bed configurations, we complemented the simulation with the cross-entropy method for optimisation. The proposed analytical method gives AH not only the ability to optimise service delivery by calculating the necessary number of beds based on existing admission records, but also a rigurous method to foresee demand for admissions that is more robust than the previous method based on crude averages.

2. SUBACUTE BED ALLOCATION

This study comprises four wards in the HRH, referred to as a, b, c and d, and two in RTRC, named x and y. The wards in HRH currently have a total capacity of 104 beds and are dedicated to geriatric evaluation



(a) Patient flows.

(b) Combined approach.

Figure 1. Representation of patient flows from origins to subacute wards (left) and combined simula-tion/optimisation approach used to solve the subacute bed allocation problem (right).

and management and to rehabilitation, whereas the wards in RTRC accomodate 42 beds, mostly dedicated to rehabilitation. Figure 1a shows a schematic of patient flows. Patients may arrive from the community, other hospitals, the Emergency Department of AH and other wards. For the purpose of this paper, the origin of the patients is irrelevant. Transfers between wards most often represent a care type change, e.g., acute to subacute, but there are also transfers between subacute wards that do not involve a care type change and also need to be captured. A small percentage of patients arriving to ward b may be transferred to a.

We performed a preliminary analysis of admission data to try to identify patterns in admission times and lengths of stay. As part of this analysis, we tried to determine if the patients' ailments, identified by AN-SNAP codes[1] were correlated to the wards the patients were admitted to. We used correspondence analysis to clarify the relation between patients' ailments and assigned wards (correspondence analysis is a method to compare data to the assumption of independence). We found that 78.8% of the variability corresponds to these two dimensions. Nevertheless, we did not use the patients' ailments directly for regression as many of the AN-SNAP codes were shared among most wards in HRH and therefore they did not perform well as explanatory variables.

3. METHODOLOGY

The approach we propose, shown in Figure 1b, combines simulation and optimisation to produce the optimal number of beds per ward given the bed demand data as follows. First, simulation is used to reproduce the patient stays in beds, using as input the appropriate inter-arrival times between patients obtained by bootstrapping actual admission data. Parametric bootstrap is used to feed the simulation process with random admission data that closely follow past and current admission trends, as described in subsection 3.1. The output of the simulation is then summarised in statistics, which are used as the criterion for optimising bed allocation. Second, optimisation, which is represented in Figure 1b as the outer loop, searches for the optimal combination of numbers of beds on each ward. With the proposed numbers of beds per ward and bootstrapped admission data, the inner simulation loop runs multiple simulations in parallel on every iteration with the aim of assessing the performance of bed arrangements. This is done by calculating performance statistics using the results of all the simulations in the current iteration of the optimisation loop. The rules that define the individual simulations are explained in subsection 3.2. If the outcome of the inner simulation loop fulfills some predefined termination criterion, the optimisation terminates and we have an optimal (or near-optimal) solution; otherwise, the summary statistics are fed back into the optimisation algorithm (in this case the cross-entropy method, whose working principle is explained in subsection 3.3) and used to direct the search for a new and improved combination of numbers of beds on each ward. The remainder of this section explains each of the components of the solution approach in more detail.

[1]The Australian National Subacute and Non-Acute Patient (AN-SNAP) classification Version 3, from http://ahsri.uow.edu.au/content/idcplg?IdcService=GET_FILE&dDocName=UOW119626&RevisionSelectionMethod= latestReleased, accessed on January 20 2015.

3.1. Input data and parametric data bootstrap

We built the list of events of the simulation using two types of data sets:

1. *Actual data*. This is the actual data set provided by the hospital, which contains information about patient arrivals and lengths of stay.
2. *Artificial data*. These data sets were calculated using parametric bootstrap. The bootstrap uses linear regressions from a previous study [García-Flores and Sparks, 2014] to simulate admission data and length of stay. The models used log-normal for first daily arrivals, logistic regression for simultaneous arrivals, Box-Cox t distribution for lengths of stay, and gamma distribution with changing parameters for day of the week and season for inter-arrival times.

Bootstrapping enables us to feed the simulator with data that captures the existing patterns of patient admissions from the original records. More specifically, the artificial data sets simulate the rules and criteria used to assign wards to patients and also reproduce the necessary transfers between wards, which are built into the regression models used to produce the data sets.

3.2. Simulation

All beds are empty at the start of the simulation. The calculation of metrics begins only after a number of simulation days have passed, in order to let the simulation process reach steady state. The patients occupy a ward according to the following logic: every admitted patient has a 'preferred' ward, that is, the ward the patient was assigned to according to the actual records, or according to the distributions obtained from the actual records and which were used to produce the bootstrapped data. This preferred ward comes from the admission records provided by the hospital. If there is a bed available in this ward, the patient is admitted there. If not, we define ward precedence as the two groups of wards for HRH and RTRC that represent the order in which the simulation looks for empty wards. For example, suppose that the preferred ward for an admission is a, which belongs to the first group of wards. The simulator will first try to find a bed in a, and if this is not possible, it will attempt to do so in b, c and d, in that order, and if it still cannot find a bed, this patient will go to the queue and keep a as a preferred ward. If the patient in the above example finds, say, a bed in d, the simulation registers this as a ward mismatch, but if the patient does not find a bed at all then he/she will join the queue and wait for a bed at a later day.

The performance of a single simulation run is assessed according to:

$$\begin{aligned} \text{performance} = &-10.00 * \text{total patient-days in queue} \\ &-100.00 * \text{total ward mismatches} \\ &+900.00 * \text{average bed occupancy} \\ &-20.00 * \text{difference from current bed use}, \end{aligned} \tag{1}$$

where the weights [-10.00, -100.00, 900.00, -20.00] can be modified according to the priorities of AH staff. The last term is justified on the grounds that there is a small cost opening and closing beds. The assumption is that the current system mostly has patients allocated to the right wards.

3.3. Optimisation

Cross-entropy estimates a sequence of parametric sampling distributions that converges to a distribution with probability mass concentrated in a region of near-optimal solutions. Consider the optimisation problem $\max_{x \in \mathscr{X}} S(x)$, having $\mathscr{X}^*$ as the set of optimal solutions. Let $\mathbf{v}$ the reference parameters of a family of probability density functions (PDF) $\mathrm{F} = \{f(\cdot;\mathbf{v}), \mathbf{v} \in \mathscr{V}\}$ on $\mathscr{X}^*$. Cross-entropy (Algorithm 1) generates a new F on each iteration, converging to $\mathscr{X}^*$ and terminating with a solution in the form of a set of degenerate distributions.

4. RESULTS AND DISCUSSION

We conducted the experiments as follows. First, actual admission data was used to calculate the optimal number of beds in all wards considering two cases, one in which ward b is included in the analysis, and one in which b is not included, in which case admissions to b will not be considered as input to the subacute optimisation problem, unless they are transferred at the end of their stay in b to another subacute ward. These scenarios

Algorithm 1 – Cross-entropy for subacute bed allocation

1: Choose an initial parameter vector $\mathbf{v}_0 = \hat{\mathbf{v}}_0$; $t \leftarrow 1$; $d \leftarrow 5$.
2: **while** $\hat{\mathbf{v}}_{t-d} \neq \hat{\mathbf{v}}_t$ **do**
3: Generate a sample $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from the density $f(\cdot; \mathbf{v}_{t-1})$.
4: Compute the sample's $(1-\rho)$ quantile $\hat{\gamma}_t$ of the performances

$$\hat{\gamma}_t = S(X_{k,\lceil (1-\rho)N \rceil}) ,$$

where $\lceil \rceil$ denotes the integer part.
5: Use the same sample to find the optimal reference parameter $\tilde{\mathbf{v}}$ by cross-entropy maximisation, using the following expression obtained from a Bernoulli distribution (see Rubinstein et al. [2013] for details):

$$v_j = \frac{\sum_{k=1}^{N} I_{\{S(X_k) \geq \hat{\gamma}_t\}} X_{kj}}{\sum_{k=1}^{N} I_{\{S(X_k) \geq \hat{\gamma}_t\}}} , \tag{2}$$

where $I_{\{S(\mathbf{X}_k) \geq \hat{\gamma}_t\}}$ is an indicator variable that takes the value 1 if condition $S(\mathbf{X}_k) \geq \hat{\gamma}_t$ is met, zero otherwise.
6: Smooth to avoid zero values of $\tilde{\mathbf{v}}$ in the first iterations: $\hat{\mathbf{v}}_t = \alpha \tilde{\mathbf{v}} + (1-\alpha)\hat{\mathbf{v}}_{t-1}$
7: $t \leftarrow t+1$.
8: **end while**
9: The solution is the single value of the degenerate distribution defined by $\hat{\mathbf{v}}$.

were requested because ward b has a different patient care mix than the other wards, despite providing the same care type. The patients have a much shorter length of stay (7 days on average), compared to 24 on average across the other wards; besides, patients are mostly admitted directly from the Emergency Department. Second, we prepared three additional data sets using the bootstrapping procedure described in subsection 3.1 and repeated the simulation four times, one for the raw data set and one for each bootstrapped data set. The global performance measure of the corresponding simulation run is in this case the average of the individual performance measures as calculated using Equation (1). When using the bootstrapped data, we also tested the optimal number of beds on every 'free' ward considering that the number of beds in all the other wards remained fixed, according to the current numbers of beds in use (i.e., the reference values). For example, if ward a is free, the code optimises the number of beds in this ward only, considering that the number of beds in the other wards is fixed to 24 in ward b, 28 in c, 28 in d, 30 in x and 12 in y.

Table 1 shows the results, obtained using a Dell PowerEdge R630 Rack Mount Server with two Intel Xeon E5-2690v3 at 2.6GHz (48 cores), coded in the *R* language[2]. These indicate that the total number of beds should increase between two and eight beds, depending on the run. When optimising the number of beds of individual wards, the results also show that the optimiser consistently recommends increasing the number of beds in one or two in most cases. Comparing the number of beds needed in winter versus the number of beds needed in summer for every year from 2010 to 2013 (not shown due to space limitations), we found that the beds needed in winter periods are greater than those needed in the summer, which is an expected result given the experience of AH staff.

5. CONCLUSIONS AND FURTHER WORK

The allocation of limited and costly health services to an ageing population whose demand for healthcare is increasing is a tremendous challenge. In this paper, we tackled the problem of calculating the optimal number of beds needed in the subacute wards of a major hospital. Our methodology combined the cross-entropy method for optimisation, the simulation of subacute ward occupation, and parametric bootstrapping of the data. CE is a modern optimisation method based on the fact that cross-entropy divergence can be used as a measure of closeness between two sampling distributions, shifting the sampling of the solution space towards a region of near-optimal solutions. The optimiser matches hospital staff intuition on the mechanics of admissions and recommends a moderate increase in the current number of beds. The model can be easily extended to consider additional information when it becomes available, e.g., actual waiting times or more

[2]Version 3.1.2, from http://cran.r-project.org/, accessed on January 8, 2015.

Table 1: Summary of results (number of beds). Columns named *All wards* refer to the optimisation of all wards simultaneously, whereas columns named *Individual wards* refer to the optimisation of a single ward, keeping the number of beds in every other ward as it currently stands. When ward *b* is not included in the set of subacute wards, it is not available (NA). Optimisation of individual wards was done only when using the bootstrapped data sets as input, but not when only using the actual admission records.

	Ward *b* included					**Ward *b* not included**				
		Actual data		**Bootstrapped**			**Actual data**		**Bootstrapped**	
Ward to optimise	**Reference**	*All wards*	*Individual wards*	*All wards*	*Individual wards*	**Reference**	*All wards*	*Individual wards*	*All wards*	*Individual wards*
a	24	34	-	18	25	24	26	-	25	25
b	24	24	-	24	25	NA	NA	NA	NA	NA
c	28	26	-	28	29	28	28	-	27	29
d	28	25	-	36	29	28	28	-	29	29
x	30	21	-	30	30	30	33	-	32	30
y	12	24	-	12	12	12	12	-	12	12
TOTAL:	146	154	-	148	150	122	127	-	125	125

accurate ward allocation rules.

Although the proposed methodology is sound, relatively fast and easy to implement, it has shortcomings. First, CE is a heuristic, and as such it does not guarantee an exact optimal solution on every run. Second, the performance of the method depends on the number of simulation runs per iteration, and this is limited by the capacity of the computer used to solve the problem. Third, there may be some bias in the distributions used for bootstrapping, the magnitude of which is documented in García-Flores and Sparks [2014]. Despite these shortcomings, the method has been shown to be practical, well behaved and to produce results that are in line with AH staff's expectations.

REFERENCES

Bachouch, R., A. Guinet, and S. Gabouj (2012). An integer linear model for hospital bed planning. *International Journal of Production Economics 140*(2), 833–843.

de Bruin, A., R. Bekker, L. van Zanten, and G. Koole (2010, July). Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research 178*(1), 23–43.

García-Flores, R. and R. Sparks (2014). Exploratory data analysis for subacute bed allocation problem. Technical Report EP149741, CSIRO Digital Productivity and Services Flagship.

Motorola (2010). Reducing healthcare costs with supply chain best practices. Technical Report IB-ISG-HCSPLYBP-0610, Motorola Inc., 1301 E. Algonquin Road, Schaumburg, Illinois 60196 USA. http://www.commenco.com/wp-content/uploads/2014/11/ApplicationBrief_Healthcare_Healthcare-Industry-Brief-0610.pdf.

Neves, L., M. Nascimento, D. Oliveira, A. Martins, M. Godoy, P. Arruda, D. de Santi Neto, and J. Machado (2014, September). Multi-scale lacunarity as an alternative to quantify and diagnose the behavior of prostate cancer. *Expert Systems with Applications 41*(11), 5017–5029.

Rosen, A., G. Gurr, and P. Fanning (2010). The future of community-centred health services in Australia: Lessons from the mental health sector. *Australian Health Review 34*(1), 106–115.

Rubinstein, R. (1997). Optimization of computer simulation models with rare events. *European Journal of Operations Research 99*(1), 89–112.

Rubinstein, R. (1999). The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability 1*(2), 127–190.

Rubinstein, R., A. Ridder, and R. Vaisman (2013). *Fast sequential Monte Carlo methods for counting and optimization*. John Wiley and Sons.

The Treasury, Australian Government (2015). 2015 Intergenerational Report Australia in 2015. Technical report, Commonwealth of Australia. http://d3v4mnyz9ontea.cloudfront.net/2015_IGR.pdf.

Wang, Y., W. Hare, L. Vertesi, and A. Rutherford (2011, May). Using simulation to model and optimize care access in relation to hospital bed count and bed distribution. *Journal of Simulation 5*(2), 101–110.

Zhang, Y., M. Puterman, M. Nelson, and D. Atkins (2012, Mar–Apr). A simulation optimization approach for long-term care capacity planning. *Operations Research 60*(2), 249–261.