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

Hospital Event Simulation Model: Arrivals to Discharge

D. Ben-Tovim [a], J. Filar [b], P. Hakendorf [c], S. Qin [b], C. Thompson [d] and D. Ward [b]

[a] *School of Medicine, Flinders University, Australia*
[b] *School of Computer Science, Engineering and Mathematics, Flinders University, Australia*
[c] *Flinders Medical Centre, Australia*
[d] *School of Medicine, Adelaide University, Australia*
Email: jerzy.filar@flinders.edu.au

Abstract: Many Australian public hospitals operate under strict resource constraints. Arguably, this is manifested in higher incidence of ambulance ramping and patient flow congestion episodes, which has led to an increase in public complaints and, possibly, sub-optimal health outcomes for patients. Consequently, there is a well accepted need to make best use of all available information and domain knowledge to ensure that hospital resources and expertise are utilised more efficiently, for the benefit of patients.

The latter is not a simple task since hospital operations involve complex interactions among many groups of health professionals utilising limited physical facilities and equipment. This is further complicated by the inherent variability of patient responses to treatments. Indeed, the stochastic nature of the demand process, as well as uncertainty in durations of medical treatments and patient recovery, lead to probabilistically distributed bed availability. Fortunately, in Australia, hospitals are "data rich" in the sense that reliable records of patient journeys have been kept for many years. While older data may reflect procedures and priorities that are no longer in place, data from recent years may be regarded as quite robust, especially in cities that have not experienced major demographic changes. Thus there is an opportunity to apply modern tools of mathematical, statistical and simulation modelling to enhance our understanding of key processes that influence a hospital's operations. The understanding so obtained can then be used to assist hospital staff in devising operational procedures that are likely to minimise disruption without adversely impacting the public service provided to the patient population.

In this paper we outline the Hospital Event Simulation Model: Arrivals to Discharge (HESMAD) to describe the patterns of patient flows within the Flinders Medical Centre, an urban teaching hospital. The logical design of HESMAD was developed through extensive consultation with colleagues from the hospital. In particular, patients within HESMAD are not modelled as identical entities, rather, they are assigned different attribute values such as mode of arrival, triage category and division to reflect the typical profile of all patients. Patients go through a set of physical units and process modules that model various physical areas, processes, interactions and behaviours within the hospital to replicate a wide spectrum of patient journeys. Hospital and patient data from 2012 to 2013 were used to fit various probability distributions, for instance the waiting times for treatment or discharges. The model allows for a realistic representation of patient flows, at a level of resolution that was deemed appropriate by the hospitals data management experts. The model has been validated against historical data and through consultation with health care and hospital experts.

Within space limitation we provide an outline and a brief discussion of HESMAD's structure, features, capabilities, design decisions and development. In addition, we provide a brief case study demonstrating the potential applicability of HESMAD for 'what if' analyses of hospital interventions. While all discussions are specific to the Flinders Medical Centre, the methodology used within HESMAD is generic enough to apply to other public hospitals in Australia.

Keywords: *Hospital operation, discrete event simulation, Poisson processes, simulation, modelling*

1 INTRODUCTION

Australian public hospitals are facing challenges due to increasing demand and budget cuts. The luxury of keeping spare capacity, namely, some empty beds, for peak demand has been replaced by the pressure to operating at, or close to, full capacity much of the time. Naturally, this translates to more frequent occurrence of congestion episodes accompanied by longer waiting times and other adverse effects.

Flinders Medical Centre (FMC), an urban teaching hospital of about 550 beds in southern Adelaide, has been a world leader since 2002 in using Lean Thinking in the hospital context as an improvement methodology to remove inefficiencies that lower the quality of patients care. That program has enabled the hospital to provide safer and more accessible care, Ben-Tovim et al. (2008). However, it has been recognised that in order to take full advantage of the Lean Thinking perspective, detailed quantitative modelling is needed to extract best information from a very rich, but complex, patient journey database (PJD).

In this paper we outline the development of the Hospital Event Simulation Model: Arrivals to Discharge (HESMAD) to describe the patterns of patient flows within FMC. The model is a result of exploiting data in PJD with the help of mathematical and statistical modeling techniques to design functions and processes that are embedded in a discrete event simulation system that supplies a convenient interface with domain experts: doctors, hospital managers and other health care professionals. Indeed, the current structure of HESMAD reflects many iterations of refinements based on feedback from these industry experts.

Importantly, patients within HESMAD are not modelled as identical entities, rather, they are assigned different attribute values such as mode of arrival, triage category and division to reflect the typical profile of all patients. Hospital and patient data from 2012 to 2013 were used to fit various probability distributions, for instance the waiting times for treatment or discharge. The model allows for a realistic representation of patient flows, at a level of resolution that was deemed appropriate by the hospital's data management experts. It has been validated against historical data and through consultation with healthcare and hospital experts.

Currently, HESMAD is being used to test hypotheses about main causes of congestion episodes, scenarios for flagging the possible onset of these episodes, and proactive operational interventions intended to mitigate against disruptions due to unanticipated increases in demand. The remainder of this paper is devoted to the description of HESMAD's structure, features and one illustration of its many capabilities.

2 STRUCTURE OF HESMAD

HESMAD is constructed using a series of interconnected components; the highest-level of HESMAD's structure is illustrated in Figure 1. The components comprising HESMAD can be broadly classified into (i) physical units and (ii) process/functional or resource modules which we simply call process modules. Physical unit represent the key treatment areas within FMC which are responsible for different aspects of patient care. A key property of the physical units is that they each have a different resource that sets its operational boundaries. On the other hand, the process modules represent the particular events/decisions/processes and control mechanisms that affect a patient's journey as they receive care within a physical unit. In terms of scale physical units reflect the macro-level design and structure of the hospital and process modules define behaviour both within and impacting physical units.

We note the absence of two key divisions from physical units, namely, Mental Health (MH) and Women's and Children (WCF). During the model's design a decision was made to focus on ED, Medicine and Surgery that are most at risk of patient bottlenecks and congestion. While WCF serves many patients, it is less prone to overcrowding and it is largely separate from the rest of FMC; utilising independent inpatient and emergency sections. Mental health patients require specialised care and hence are rarely treated outside of the MH. However, modelling their contribution to boarding times is crucial as they do impart significant stress on ED where they can occupy emergency spaces for extended periods. Consequently, their boarding time is explicitly modelled within the boarding module using empirical length of stay (LOS) distribution drawn from the PJD.

The main state variable we are modelling is occupancy on any given day, τ, denoted by $X(\tau)$. Conceptually, the model which computes $X(\tau)$ reported in Figure 2 is a balance equation

$$X(\tau) = f(X(\tau - 1), A(\tau), D(\tau)), \tag{1}$$

where $A(\tau)$ and $D(\tau)$ are arrivals and discharges on day τ, and are realisations of the corresponding stochastic processes that are modelled in their individual process modules. Due to space constraints we are unable to spell out the multi-component structure of the balance equation 1. The latter will be supplied in a full journal length

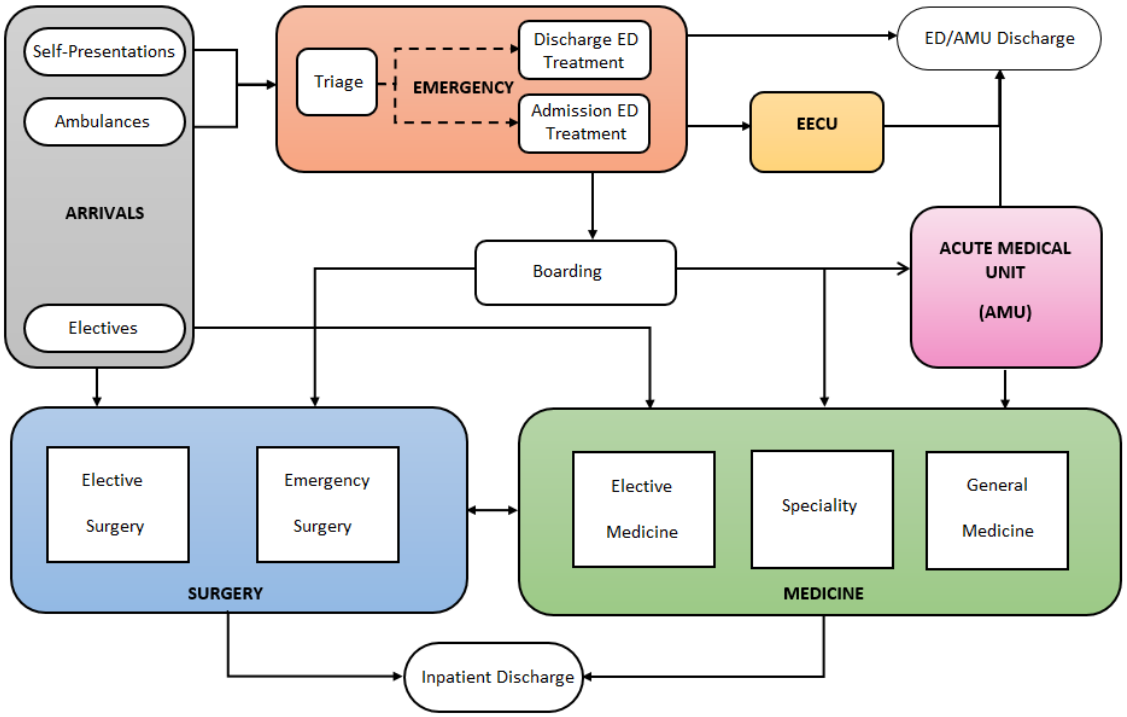

Figure 1. High-level structure of HESMAD

article that is currently in preparation. However, we illustrate one process module and one physical unit in more detail in order to allow the reader an insight into the mathematical constructs embedded in HESMAD. A summary of the key modules and units is also provided in Table 1 and Table 2.

2.1 ED Physical Unit

The ED module(see Table 1) simulates FMC's emergency department which treats all arrivals 24 hours a day, 7 days a week. All non-elective patients in the arrival module (see Section 2.2) are processed by the ED module. This physical unit comprises four types of process modules: Triage, ED Buffer, ED Streams and Resources. Triage score is used as an indicator of severity of a patient's condition/symptoms/cormabidity and hence determines queuing priority, treatment stream and treatment time in ED. In HESMAD we regard it as a realisation of a random variable variable C whose conditional probability distribution (depending on the mode of arrival, namely, ambulance versus self-presentation) is approximated by

$$\hat{P}\left(C = c | M = m\right) = \frac{n_{c,m}}{n}, \quad c \in \{1, 2, \cdots, 6\} \quad \& \quad m \in \{0, 1\}, \tag{2}$$

where C is triage score and M is mode of arrival. The values n and $n_{c,m}$ represent the observed number of ED arrivals, and ED arrivals of type m and triage category c, respectively. As well as assigning a triage score, triage nurses also assign patients to one of two emergency streams based on their likelihood of admission or discharge: s_a or s_d streams. Discussions with experts suggest that, in most cases, the nurse makes a correct allocation to the s_a or s_d streams. As data relating to the accuracy of the assignment were unavailable, the actual admission rate for each triage score was used to determine treatment stream. Namely, the conditional probability of being assigned to a given stream is given by

$$\hat{P}\left(S = s | C = c\right) = \frac{n_{s,c}}{n_c} \approx \frac{n_{a,c}}{n_c}, \quad s \in \{s_a, s_d\}, \quad a \in \{0, 1\}, \tag{3}$$

where S is the ED stream and n_c is the observed number of ED arrivals with triage category c. Since the number $n_{s,c}$ of patients in triage category c who were initially allocated to stream s is unavailable, we approximated it by $n_{a,c}$ the observed number of ED arrivals with triage category c, and admission status a, respectively.

After assignment, patients either proceed directly for treatment in an s_a or s_d stream (if a resource is available), or enter the ED buffer. Patients in the ED buffer are queued based on their triage score, with category 1

Table 1. HESMAD physical units and descriptions

Physical Unit	Description
ED	Emergency Department (36 treatment spaces): models initial acute care of patients presenting without appointment; either as self presentation or by ambulance (see Section 2.1).
EECU	Extended emergency care unit (8 beds): fast-turnover inpatient ward designed for patients requiring up to 24 hours of monitoring within the ED before being sent home.
AMU	Acute medical unit (30 beds): assessment unit for general medical and acute care elderly patients. AMU provides additional assessment (beyond ED) that sorts and allocates patients to other inpatient teams, introduced to reduce waiting times in ED. AMU patients are either sent for long term medical care, or require only short admission and are discharged within 24–48 hours.
Medicine	Largest inpatient division within FMC (182 beds). Treats patients requiring emergency hospital care for medical conditions with complex and multi system disease. Three treatment streams are modelled based on how patients enter the division: speciality - referred directly from ED, general medicine - patients underwent AMU assessment and, elective - non-emergency arrival undergoing non-surgical elective procedure requiring overnight care.
Surgical	Other primary inpatient division (142 beds) that treats patients requiring surgery. Two treatment streams based on mode of entry: emergency surgery and elective surgery.

Table 2. HESMAD process modules and descriptions

Process Module	Description
Arrivals	Generates emergency and elective patients within the simulation (see Section 2.2).
Triage	Emergency arrival priority system for sequencing treatment. Empirical distributions based on mode of arrival.
ED Buffer	Post-triage patient queue. Order of service is based on assigned triage score, priority given to triage 1 and 2 patients.
ED Streams	Assesses, treats and/or stabilises emergency patients. Contains two, physically separated, streams based on likelihood of admission for inpatient care. Estimated treatment time is drawn from empirical distributions based on ED stream, triage score and mode of arrival.
Dispatch	Allocates patients to modelled inpatient streams. Exploits probabilistic distributions influenced by triage score.
Boarding	Post ED treatment queue for admitted patients waiting for an inpatient bed. Boarding patients occupy emergency treatment space, reducing ED throughput.
Treatment	Simulates a patient's period of treatment associated with their current inpatient stream. LOS is drawn from empirical distributions for relevant stream.
Resources	Finite resource utilised by patients within physical units. Resources includes ED treatment spaces and inpatients beds. Restricts patient flow between physical units based on availability.
Operating hours	Schedule module which controls inpatient operation hours relevant to various patient streams.
Discharge	Controls discharge processes for patients exiting FMC from a physical unit. Includes administrative delay relating to processing a discharge as well as a cleaning delay on freeing up a resource.

and 2 patients having priority. Priority 1 patients are allowed to bypass queuing as they represent the most urgent cases requiring immediate medical attention. If their allocated stream's resource is unavailable, they either seize a resource from the other stream (borrowing) or proceed without one (bypassing). This recreates flexibility observed within ED in terms of its capacity and within the PJD where all category 1 arrivals have zero waiting time. Once a patient seizes an ED resource they proceed for treatment within the ED stream module. The LOS for their treatment is based on the stream and the patient's priority and mode of arrival. The LOS is modelled using empirical LOS distributions $\hat{F}_{c,m,s}(t)$. Empirical distributions are used for two reasons, firstly the PJD provides sufficient data to do so and, secondly, routine parametric distributions do not adequately model the observed data in terms of goodness-of-fit (e.g., bootstrap Komologorov-Smirnov test, Conover (1999)). After treatment, patients depart the ED module for either the discharge or dispatch modules.

2.2 Arrival Process Module

As the name suggests the arrival module is responsible for generating an inflow of patients into the simulation in a manner that is consistent with the behaviour of arrivals observed at FMC. Two, distinct, arrival processes are considered: emergency (*unscheduled*) and elective (*scheduled*), with the emergency being further subdivided into ambulance (amb) and self-presentations (sp). Due to the predominantly unrelated nature of emergency arrivals, they are typically modelled by memoryless Poisson processes. Furthermore, the emergency arrival rates of these processes differ across the day and exhibit strong trends across weekdays. The emergency arrivals

$$\{N_j^m(t)\}_{t\geq 0}, \quad j \in \{1, \cdots, 7\}, \quad \& \quad m \in \{0, 1\}, \tag{4}$$

are modelled using inhomogeneous Poisson processes with rate functions $\lambda_j^m(t)$ that depends on the day of week j and the mode of arrival m. Each of these λ_j^m functions are approximated by a piecewise-constant function

$$\hat{\lambda_j^m}(t) = L_{j,h}^m, \tag{5}$$

where L is a fitted constant value for $t \in [h, h+1]$, with $h = 0, \cdots, 23$.

While a detailed discussion of the validity of fitting inhomogeneous Poisson processes is not possible within page constraints of this paper, they have been validated both numerically and statistically using the PJD data. The latter validation relied on techniques recommended in Kim and Whitt (2014), as straightforward statistical testing involving goodness-of-fit often leads to spurious conclusions resulting from issues of over-dispersion, and rounding. In particular rounding proves to be a significant issue due to the use of triage time which is both rounded (to the minute) and becomes less correlated to arrival time during busy periods of ED.

Due to limited bed capacity within the hospital the number of electives performed day-to-day can vary significantly based on the hospital's occupancy. Through delaying electives the hospital is able to exert some control over its workload and can increase or decrease the number of electives at times of low or high occupancy. To capture this behaviour the number of electives on a given day of the week is first generated as a coarse estimate (based on relevant averages) and is then adjusted based on the hospital's occupancy level. A range is specified for this adjustment, based on observed minimum and maximum arrivals with the PJD. Any deferred arrivals are backlogged with their arrival occurring at the next feasible time slot. The inclusion of minimum arrival criteria results in the simulation being able to exceed nominal admission capacity (which does happen in practice). In such a case the throughput of ED is impacted - in terms of patients' extended boarding - and only returns to normal operation after occupancy drops below capacity again.

2.3 AGILE model development

Developing HESMAD involves planning, analysis, design, implementation as well as validation and verification. The planning, initial analysis and design phases began well in advance, during the preparation of the peer reviewed ARC Linkage proposal, and the analysis and design phases continued in response to regular feedback from industry partners in the project. We utilised the so called Agile development approach. An Agile approach usually applies time based iterative and evolutionary development cycles that allow adaptive planning, incremental delivery and rapid and flexible response to change. Short, iterative and incremental cycles of development took place as the output of each HESMAD development cycle was compared with historical data and presented to members of our multi-disciplinary team, especially our colleagues from FMC. This encouraged frequent communication that played a vital role in improving functionalities and usefulness of HESMAD.

The model was implemented in Anylogic (e.g., see http://www.anylogic.com). AnyLogic provided a flexible modelling platform supporting multi-method modelling and customisability, permitting both quick construction and modification. However, none of the particular design choices in any way limit the simulation from being implemented in other modelling packages which offer similar capabilities.

3 VERIFICATION AND VALIDATION

A crucial yet difficult step in developing any simulation model is its verification and validation. In particular a fine balance is needed between developing a simulation that is an accurate representation of the phenomenon being modelled but not so over parameterised that it involves assumptions which are unsubstantiated and cannot be modified from its base case. In the case of HESMAD, the simulation needed to (i) provide reasonable approximation of the macro level flow of patients through FMC from arrival to discharge and (ii) be general enough to allow for the analysis and testing of potential interventions and structural alterations. The verification and validation process is ongoing while the simulation is being refined and further developed. Some of the techniques that have been used to validate/verify HESMAD include:

- Historical data validation. Under current operating procedure scenario the simulation behaviour was tested at different levels of granularity (across single, multiple or all physical units). This was done through comparison of simulated and historical data metrics such as midnight occupancy, waiting times, total ED LOS or total hospital LOS.
- Consultation: relates to agile model development, see Section 2.3. Iterative consultation with expert team members to (a) verify the simulation implementation matches the conceptualised design and (b)

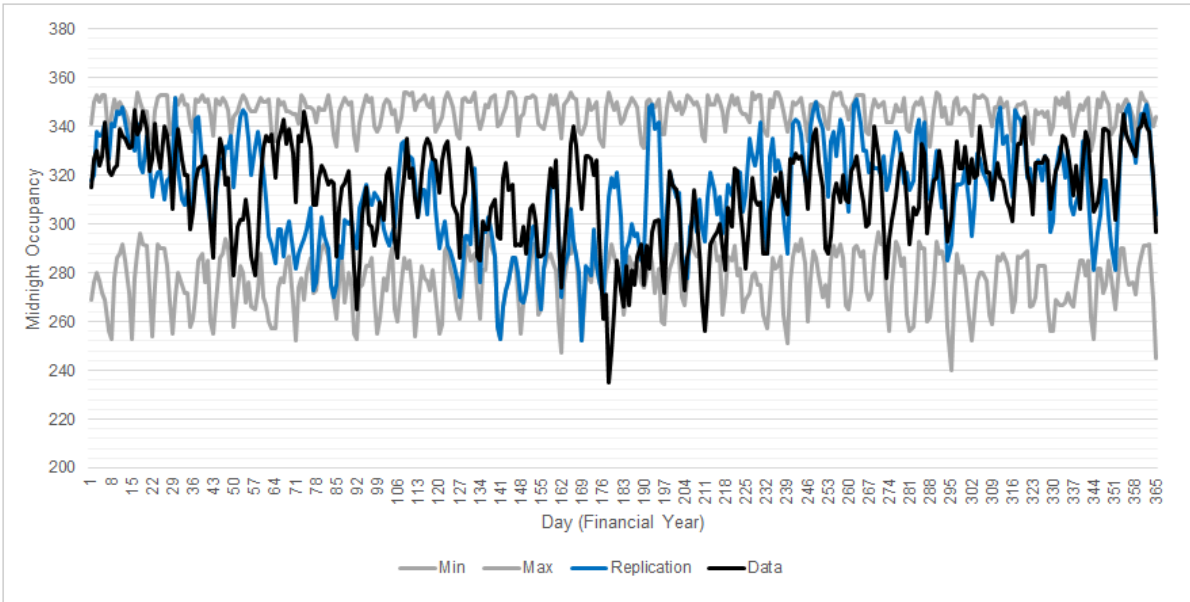

Figure 2. Midnight occupancy data, simulated vs actual midnight occupancy for 2012-2013 financial year.

provide face validity to the simulation. Recently, additional consultation with health-care experts outside of the primary collaboration group has also been used to provide an additional level of validation.

- Modular and unit validation/testing: input-output testing of individual process modules, including statistical testing of their outputs and assumptions, such as fitted distribution, for instance, where inhomogeneous Poisson processes were used (see Section 2.2).
- Extensive debugging of simulation code to reduce likelihood of implementation errors.

An example of the last type data of validation is presented in Figure 2 whereby midnight occupancy data for the 2012-2013 financial year (in black) is plotted against synthetic midnight occupancy values generated using HESMAD (for 100 replications). As can be seen, while there are differences, trend wise the simulation and historical closely resembles each other, even for a single simulation replication (in blue). The simulated occupancy has the same prominent weekly trend associated with hospitals and observed in the actual data, both have similar means, approximately 315, and the real data is well bounded by maximum and minimum recorded occupancy values for each day (in gray). A slight discrepancy does occur during the Christmas period which is not adjusted for explicitly within the simulation, however, it can also be seen in the single realisation case that low occupancy events are permissible and, indeed, do occur.

4 MODEL APPLICATION

We present a sample application of HESMAD illustrating how it can be applied to evaluate the impact of potential discharge strategies (interventions), including the current practice, on the hospital's occupancy.

The scenarios considered are summarised as follows:

1. Base case. Standard Hospital Operation Scenario (no interventions).
2. A 24 hour discharge option; outside of schedule.
3. Accelerated discharge option. No administrative delay on inpatient discharge.
4. Discharge patients with LOS $>$ 21 days (prevent long stay outliers).
5. Reduce LOS by half a day for all patients with LOS of at least 2 nights.

Actual data is available only for the base case (Scenario 1). The remaining interventions scenarios were formulated in consultation with our industry partners and indicated their interest in capturing their likely impact. The results outlined below were in concordance with their professional intuition. Of course, one goal of HESMAD is to perform tests of "what if" questions. Whereas using simple average estimates may give some indications of impact of interventions, HESMAD demonstrates much more detailed impacts such as where queues are building up, frequency of extreme events, which physical units are idle, during which periods and

Table 3. Summary table of discharge/intervention scenario experiments (100 replications). All values except standard deviation are are rounded to nearest integer.

Metric	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Mean Occupancy	315	311	317	290	305
Mean Std Dev	16.5	16.9	17.1	17.4	18.1
Maximum Occupancy	354	354	354	351	354
Minimum Occupancy	240	230	236	218	230
Mean Red Days	9	4	15	< 1	2
Mean Amber Days	227	198	234	48	156
Mean Green Days	129	163	116	316	207

so on. Hence, simulations are essential in a preliminary phase before piloting any new discharge procedures. Space constraints prevent us from reporting a full range of these impacts.

Each realisation of the simulation (called an experiment) simulates a single financial year of hospital operation under the proposed intervention. Midnight occupancy was captured throughout the simulation and outputted for analysis. To account for variability across realisations (due to stochastic nature of the simulations) each experiment was replicated 100 times to analyse the system's average behaviour and extremes. The average occupancy profile for each scenario is presented in terms of a "stop light system" used by the SA Health. This indicates the status of the hospital occupancy where green is safe, amber is at risk, red is near operational capacity. The thresholds used for transition between status level is based on the guidelines used by SA Health (see, http://www.sahealth.sa.gov.au). The results are presented in Table 3. Note that the first four rows corresponding to occupancy metrics are measured in number of occupied beds, while the last three rows are measured in terms of units of days when the hospital is in one of the three aforementioned states.

While only an illustration, the results presented in Table 3 demonstrate an interesting and wide range of responses. Some key observations are:

- Base case provides a good match with historical data both on the basis of statistical fits and industry partners intuition.
- All but one intervention scenarios results in average occupancy reductions.
- Reducing average occupancy need not decrease variability. For instance see mean and standard deviation of occupancy under Scenario 5.
- Targeted strategies may lead to similar improvements (occupancy-wise) to global strategies. For instance, long stay outlier (> 21 days) patients in Scenario 4 comprise less than 3% of all emergency admissions but this Scenario generates significant improvements in mean occupancy < 300.
- Changes at the end of patient journeys have influence on earlier processes. For instance, this is indicated in Scenario 3 where mean occupancy actually increased despite greater flexibility in discharge times. Factors relating to elective arrivals and emergency behaviour (not presented) are being impacted due to larger intake of elective patients altering the system's base behaviour. A more in depth analysis of the influence of elective patient intake and case mix within the hospital is part of ongoing research to better understand the underlying processes.

ACKNOWLEDGEMENT

This work was supported by the ARC linkage grant LP130100323, jointly awarded to Flinders University, the Southern Adelaide Health Service (Flinders Medical Centre) and the Central Adelaide Local Health Network (Royal Adelaide Hospital). The authors acknowledge useful discussions with Dr M. Mackay.

REFERENCES

Ben-Tovim, D. I., J. E. Bassham, D. M. Bennett, M. L. Dougherty, M. A. Martin, S. J. O'Neill, J. L. Sincock, and M. G. Szwarcbord (2008). Redesigning care at the flinders medical centre: clinical process redesign using "lean thinking". *Medical Journal of Australia 188*(6), 27–31.

Ben-Tovim, D. I., M. L. Dougherty, T. J. O'Connell, and K. M. McGrath (2008). Patient journeys: the process of clinical redesign. *Medical Journal of Australia 188*(6), 14–17.

Conover, W. (1999). *Practical Nonparametric Statistics*. John Wiley and Sons, New York.

Kim, S.-H. and W. Whitt (2014). Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management 16*(3), 464–480.