USING SIMULATION TO HELP HOSPITALS REDUCE EMERGENCY DEPARTMENT WAITING TIMES: EXAMPLES AND IMPACT

Thomas Monks
Rudabeh Meskarian

University of Southampton
University Road
Southampton, SO17 1BJ, UK

ABSTRACT

In recent years, all acute hospitals in the UK have experienced unprecedented emergency department waiting times and hospital bed pressures. The consequences are overcrowded emergency departments, ambulance shortages, cancelled elective operations, low staff morale and financial penalties. To deal with the increasing numbers of patient admissions and delayed discharges hospitals must turn now to modelling and simulation to help increase their flexibility and ability to deal with demand variation. Hospitals face several issues that reduce their flexibility including the need for extreme value-for-money and specialization of care. This talk presents three ED case studies undertaken by an analytics team in the UK. The paper considers the impact of the work and challenges arising from their experiences of simulation modelling in acute hospitals. Final thoughts consider the future of ED simulation.

1 INTRODUCTION

Emergency department (ED) overcrowding is common in the United Kingdom’s (UK) National Health Service (NHS). In England, performance of EDs are publicly reported using the (infamous) four hour waiting (cycle) time target. The target was originally set to a 100% level of compliance, but was quickly relaxed to 98% in 2004 and then 95% from 2010. The pressure felt in an NHS ED is closely linked to bed pressure within the main hospital (Cooke et al. 2004). Bed pressures and subsequent poor ED performance are due to systemic issues ranging from capacity in community based care, such as social services and nursing homes, lack of investment in public health, and the aging population. While the long term solutions to the systemic problem are sought, hospitals must find ways to increase their short term flexibility to deal with variation in demand and length of patient stay. Many of the flexibility issues experienced by hospitals can be conceptualized as queuing problems. Modelling and Simulation (M&S) therefore has a key role to play in finding solutions that ease the pressure on ED’s and hospitals.

M&S has been used extensively to study EDs (Gul and Guneri 2015). The majority of studies take a discrete-event simulation approach (DES; e.g. Coats and Michalis 2001; Günal and Pidd 2009; Paul and Lin 2012; Chavis et al. 2016; Oh et al. 2016; Wong et al. 2016; Yang et al. 2016). DES studies have typically focused on the number and scheduling of resources within the ED (Saghafian et al. 2015). Few studies appear to have worked with hospitals to examine the flexibility of the ED and associated processes to deal with variation in demand. With the exception of a few large centers in the UK, hospitals are unlikely to have the finance for substantial increases in their ED staffing. Hence the efficiency and flexibility of processes has become ever more important.

This paper first presents some of the challenges UK hospitals face in managing demand and the options they have used in order to have the flexibility to meet waiting time targets in the face of hospital bed pressures. We describe available data sources for M&S and detail three case studies that supported on the ground decision making about process changes in order to improve ED performance. We then discuss the impact and use of the case studies. The final section considers future research directions for ED M&S.
2 HOSPITAL FLEXIBILITY: CHALLENGES AND OPTIONS

Hospitals face a number of challenges in managing the emergency and elective demand on their services. Five key issues that affect bed pressures and ED performance are value-for-money, medical specialization; the requirement to manage genders separately; delayed transfers of care; and facility layout.

As the NHS is publicly funded there is a general perception that any degree of idleness in resources is wasted money. Value-for-money is therefore of high importance. This is problematic because demand for services can be highly variable and acute hospitals can vary dramatically in size and hence will not be able to achieve the same waiting times for a given level of occupancy.

Medical Specialization is the division of patient care into specific disease groups and sub-groups. The importance of specialties and their sub-specialities is that it reduces the variation in the patients that doctors see and improves the quality of patient care. A secondary consequence that challenges flexibility is that specialization creates a large number of separate queues within a hospital, because specialized care is provided by ward.

The UK government has mandated that all NHS funded care must provide single sex accommodation. As with specialization this creates additional queues for inpatients beds. For example, a male patient with a neurological condition may not be able to be placed on a neurology ward that has a free bed, as the free bed is within a female designated bay.

Delayed transfers of care occur when a patient experiences a delay in their discharge home or transfer to another care provider. These delays can use up a substantial number of beds with figures of 150-200 out of 800 beds not being untypical for any given day.

Many hospitals in the UK were designed decades ago when the population and its needs were very different. This means that much of the space and layout in a hospital is not optimized for efficient patient flow or waste reduction. For example, CT scanners located 10 minutes’ walk from an ED; or surgical recovery wards located on a different floor from operating theatres. All of these delays appear minor, but typically introduce further delays. As an example, patient transfer to a recovery ward will need a porter and a nurse. This will remove a nurse from a ward for a short time which will delay ward processes.

Given these challenges, Table 1 summarizes common approaches that hospitals use in order to increase their flexibility to deal with variation in demand. It should be noted that some of these options provide short terms gains in flexibility at the expense of longer term capacity. For example, outlying of medical patients to a different specialty provides short term capacity in an ED or admission unit, but adds potential lengthy transfer procedures and increases the need for doctors to leave their usual ward(s) to make ‘safari’ rounds to different specialties. The remainder of this paper discusses three case studies that analyzed some of these approaches: pooling servers; buffers and carve out.

<table>
<thead>
<tr>
<th>Flexibility theme</th>
<th>Implementation options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buffers</td>
<td>• Admission units and clinical decision making units</td>
</tr>
<tr>
<td></td>
<td>• Discharge lounges</td>
</tr>
<tr>
<td>Pooling and specialization of resources</td>
<td>• Carve out capacity and specialization of treatment;</td>
</tr>
<tr>
<td></td>
<td>• Co-location and pooling of specialist beds;</td>
</tr>
<tr>
<td></td>
<td>• Carve out a proportion of ward capacity to act as ‘short stay’ admission beds;</td>
</tr>
<tr>
<td>Flexible workforce</td>
<td>• Advanced Nurse Practitioners (generic servers that can fulfill nurse’s role and aspects of a full doctors);</td>
</tr>
<tr>
<td>Space and capacity management</td>
<td>• Organize wards into bays of beds that can be switched between the genders;</td>
</tr>
<tr>
<td></td>
<td>• Outlying medical patients to different specialties and surgical wards;</td>
</tr>
<tr>
<td></td>
<td>• Targets for early ward rounds and discharges;</td>
</tr>
<tr>
<td></td>
<td>• Chair based emergency care for ambulatory patients;</td>
</tr>
<tr>
<td></td>
<td>• Early discharge of patients along with home care provided by the hospital;</td>
</tr>
</tbody>
</table>
3 DATA SOURCES FOR ED MODELLING

ED’s in the UK are performance managed in regards to patient time in the department. As such, ED database systems contain high quality information on patient arrival and departures times. This is accompanied by basic demographic information (age, gender), mode of arrival (ambulance or walk-in) and presenting complaint. Many process times needed for M&S will be omitted and will require primary data collection or assumptions. A more variable finding is electronic tracking of patient movement through the ED; for example, the allocation of patients to bays and use of diagnostic imaging. We have worked with systems that contain detailed tracking and those that contain none. In the latter we have used a combination of expert judgement and primary data collection. For broader studies of ED, data can be linked to a Patient Administration Systems (PAS). A PAS tracks patient transfers between inpatient wards and also provide primary and secondary diagnoses in the form of an ICD-10 code. Data linkage can sometimes be achieved using a unique patient identifier such NHS number. In some cases fuzzy matching to achieve linkage may be needed e.g. using a date and time range and demographic information.

4 CASE STUDY 1: EDUCATION ABOUT QUEUEING

Here we illustrate some simple educational tools that we used as an initial intervention in a hospital facing a severe ED waiting time problem. The work is in part inspired by SimLean educate (Robinson et al. 2014), although we developed our own models. The work focused on talking to management about flexibility and how different configurations of their processes would improve or reduce it in different scenarios. We provide three examples, extreme value for money, specialization versus pooling and thinking systemically. Validation and verification (V&V) of results was conducted by a comparison to expectations from queuing theory.

4.1 Extreme Value for Money

Within a hospital, closure of beds is often perceived to save money. Management and clinician’s clash over the need for (extreme) value for money from beds, i.e. aiming for ward occupancy (bed utilization) in the middle to upper 90 percentages, whilst needing to keep the system moving. Although the trade-off between queuing time and utilization is well known to specialists it is not well known within health care management (there is also some evidence of this within manufacturing; Suri (1998)). In this case we found that it was difficult for management to accept that, in order to deal with demand and process variation, some of the beds need to be empty for some of the time. Our initial approach to working with the hospital was to provide educational material about queuing. Figure 1a illustrates our default starting point: the trade-off between average waiting time and resource utilization. We combined the result with a simple single ward model implemented in Simul8 Professional (Simul8 Corporation 2017). The lesson aims to develop management’s understanding about the need to consider target waiting times and, given current processes, the traffic intensity that achieves it.

![Figure 1a: Classic queuing utilization trade-off.](image1.png)

![Figure 1b: Specialization versus pooling.](image2.png)
4.2 Specialization versus Pooling

The hospital faced significant delays in admitting emergency patients to an inpatient bed. That is medical patients queued on trolleys in the ED until a bed became available in an inpatient ward. This was a major bottleneck in the ED. The hospital’s management were considering splitting a single admission ward into multiple specialized admission wards, as they believed this would speed up processing and reduce queues in the ED. Figure 1b illustrates the simple model used to initially talk to management about the flexibility benefits offered by pooling resources versus the potential benefit of reducing LoS from specialization. In this example, a single ward is split into three sub-specialisms. An approximate 60% reduction in average Length of Stay (LoS) (and Standard Deviation; Stdev) was needed to deliver the same average queuing time. The actual reduction needed will vary depending on context, but the result was enough to highlight to management that it should not be assumed that a small reduction in LoS will be sufficient to improve performance. In this case senior doctors argued that it was unlikely that sub-specialization projects would substantially reduce variation or average duration in LoS. This led to a change of focus in the project.

4.3 Thinking systemically

Emergency patients that required an inpatient stay were first processed in an Acute Medical Unit (AMU). The AMU initiates a more detailed investigation of a patient’s condition and may discharge a patient home or refer them to medical sub-speciality. Ideally patients spend no more than 24 hours in an AMU. In this case, patients faced a lengthy ED ‘trolley wait’ for a bed in the AMU. The AMU was frequently blamed for its processing delays and lack of capacity. Discussions with the lead doctor for the AMU made it clear that the AMU faced significant issues in transferring patients to inpatient wards on a timescale dramatically different from the ED’s needs (e.g. 15-24 hours). As such the educational model in Figure 2a was conceptualized. The model contained three activities, resource sets and associated queues: ED assessment, AMU processing and a stay in a ward. The ward splits patients into two groups: long stays (30% in the control scenario) and short stays. The model was calibrated so that the wards run at a 98% occupancy – which was similar to what the hospital currently experienced. The results displayed in Figures 2b and 2c illustrate the message of the model. That is, a change in ward processing had a significant impact on ED overcrowding even though it was two steps removed from the ED. In this case, a very small shift in the population (1-2%) toward short stay had more impact on ED overcrowding than increasing the AMUs bed capacity by 20 beds (another ward). This was because the hospital outflow was the major constraint.
5 CASE STUDY 2: INTRODUCING A BUFFER

A common approach to tackling excessive ED waiting times is to create a ward that acts as buffer between the main hospital and the ED. These wards are often called clinical decision-making units (CDU). They offer more time to assess a patient’s treatment needs off the clock i.e. they do not count towards the NHS ED waiting time target. This comes at a cost as patients will often spend longer in a CDU than in an ED. This is in part because the same senior doctors and nurses that staff the ED must also work in the CDU, but also because attention then must turn to dealing with new patients that arrive at the front door of the hospital.

Figure 3 illustrates a simple DES model, implemented in Simul8 (Simul8 Corporation 2017), that was used by a large hospital in the UK. The objective was to identify the size of CDU (no. beds) needed to achieve the four-hour waiting time target. The scope of the modelling included the hospital’s Acute Medical Unit (AMU) that admits medical patients from the ED. In Figure 3, the rectangular boxes represent processes, for example assessment and treatment in the ED. The partitioned rectangles represent queues, for example patient waiting for admission to the AMU. Entities within the model (patients) arrive to the model with a variable inter-arrival time (IAT). The IAT varies across three factors: the patient’s mode of arrival (walk-in or via ambulance); weekday or weekend; and by hour of the day. For the hour of the day a non-homogenous Poisson process was simulated using thinning (Lewis and Shedler 1979). Entities have a variable LoS distribution in the ED depending on their classification as a minor or major emergency. V&V of model parameters and CDU logic was conducted with input from acute medical physicians. The CDU here was indeed being considered as an overflow buffer. In this case the proposed design was to admit patients to the CDU when 30 minutes or less was remaining until they breached the four-hour waiting time target. Following processing in the CDU patients are either admitted to an inpatient bed, AMU or discharged home.

![Figure 3: Clinical decision making unit model.](image-url)
Figure 4 illustrates the capacity requirements of a CDU if it were used as a buffer to achieve the waiting time target. The results are shown by varying mean process time for the CDU (an unknown parameter) between 2 and 7 hours. The hospital believed the most likely processing time to be between 2 and 4 hours. Even at 2 hours this requires a large ward (30 beds).

![Figure 4: Bed capacity requirements to achieve waiting time target.](image)

6 CASE STUDY 3: WHEN GENERIC SERVERS ARE INFEASIBLE

Figure 5 illustrates the logic of a DES model, implemented in Simul8 (Simul8 Corporation 2017), of the full acute medical pathway of a hospital in the UK. Its focus is on unblocking the outflow of an ED, as this was found to be the biggest constraint on ED waiting times. V&V of parameters and logic was undertaken by clinicians whilst black box performance was compared to historical data.

![Figure 5: High-level logic for the DES model of a medical pathway.](image)

We modelled hourly arrivals as a non-homogenous Poisson process (weekday mean = 92, weekend mean = 84). The majority of patients wait on a trolley within the ED and are admitted to the AMU. A minority of patients need to be admitted directly to a long stay inpatient ward or if they can walk to Ambulatory Emergency Care (AEC). During the day, the treatment needs of patients in the AMU are assessed by senior acute physicians. If required patients are then transferred to a specialist inpatient ward.
or, if no further acute care is needed, discharged from the hospital. All patients admitted overnight to the AMU are ‘cohorted’ until the morning when ward rounds begin again. Patients ideally spend time on the appropriate specialty ward e.g. respiratory or medicine for older people, rehabilitation and stroke (MOPRS), but may end up on a different medical specialty or a surgical ward if space cannot be found.

LoS in specialty wards can be substantial; for example, here the average LoS for MOPRS is 15 days (stdev = 19 days). Within the population there are individuals who would benefit from more intense treatment and attention; i.e. they could be processed and discharged within a shorter time frame. Identification of these patients takes place within the AMU. One option therefore is to then ‘badge’ these patients as priority individuals and distribute them amongst the general medical population. Pooling beds and staff in this manner appeared to be intuitive. However, the senior management of the hospital stated that, due to the culture and processes of long stay wards, at best, treating all staff as generic servers would not be sustainable and patient LoS targets would not be met (no data were provided). Management therefore wanted to investigate carving out capacity to act as a dedicated 72-hour intensive treatment ward.

One issue with using carve out in this manner is that medics must make predictions about which patients would benefit from more intensive treatment. The reality is that this is very difficult. The hospital was given the recommended target of 65% accuracy. That is, at any one time around 65% of patients on the ward are patients who will be discharged within 72 hours of admission.

Figure 6 illustrates the results from a scenario that reconfigured capacity so that two wards (60 beds) were carved out as a 72-hour short stay ward. Four key outputs (averages) are displayed: ED admission waiting times (hrs); medical outliers (patients who are admitted to non-medical ward); the occupancy (utilization) of the short stay ward; and the queuing time from AMU to the short stay ward. These outputs are varied by the prediction accuracy in LoS. The occupancy of the short stay ward remains near 100% until accuracy reaches 80%. This then reduces the queue from the AMU and begins to unblock the ED outflow: reducing admission waiting times. A target of 65% accuracy would be insufficient with this level of carve out.

![Figure 6: Carve out and accuracy results.](image-url)
7 IMPACT OF THE WORK

The educational models we include in case study one are well known results from M&S. Our purpose is not to argue the novelty of the results we present, but instead to argue that successful M&S studies in hospitals often require a simple starting point. The basic concepts of flexibility that we highlight are often hidden in complex models of hospital processes and hence it is difficult for management to understand why results are different from their expectations. The approach we take has much in common with system archetypes in System Dynamics (Wolstenholme 2003) and SimLean Educate (Robinson et al. 2014). An ideal study would follow-up on these initial models with a more detailed model of the competing designs, for example to estimate the actual reduction in LoS and variation that specialized wards must achieve to meet ED admission waiting time targets. However, the political pressures on hospitals often mean that decisions are made quickly and operational changes can take place over night. This is problematic given that a detailed M&S study typically takes 3-6 months (Hoad et al. 2015). In case study one, we had a single day to influence the decisions at the hospital. Hence the purpose of the study was simply to review some basic dynamics of queuing to inform the debate. The models were still influential and convinced managers to focus on the flexibility of process and capacity further downstream. A pitfall of the educational approach is that managers on the periphery of the study can mistakenly take the numbers, as opposed to the concepts illustrated, literally. Particularly as results may be presented or reused long after the M&S team have moved on. As such it is recommended that educational material is carefully (and permanently) badged and caveated. An unexpected benefit of working with the hospital in this manner is that one of the clinical leaders involved learnt a lot about the basics of M&S. The individual went on to build a similar model to Figure 2a in the System Dynamics package Stella Professional (ISEE Systems 2017). This model went on to be influential in further persuading the Chief Operating Officer about where to target initiatives to reduce ED overcrowding.

The second case study was again influential in hospital decision making. It is unusual in the sense that its impact was to rule out an option as feasible. In this case management chose not to implement a CDU. As such it saved time and effort in implementing a new process that would lack the capacity to solve the overcrowding problem. The timelines for project were only a matter of days, but took place during a larger study of the ED’s efficiency where parameters and basic logic had been validated by clinicians. Hence the modeler had access to preprocessed, accurate and detailed time dependent inter-arrival time data along with existing estimates of patient routing and processing times. The impact here was simply being in the right place at the right time.

The final case study was a detailed model of an acute medical pathway and associated bed blocking. There were two key impacts of the study. Firstly, the hospital re-assessed the accuracy target that had been initially agreed and adopted a more stringent policy. Indeed, the management team recognized that outlying non-short stay patients to the ward would have an extremely detrimental effect on performance, due to subsequent transfer delays. The results also prompted a re-cap of the educational models explaining the trade-off between queuing time and occupancy. The results indicated that an average occupancy above 90% was not feasible; however, this was common in the hospital with nearly all in-patient wards at 95%-98%. Secondly, the need to select the correct patients, those that would benefit from intensive treatment, meant that an investigation of tools to help predict a patients length of stay was also needed (e.g. Carter and Potts 2014; Juma et al. 2016).

Our experience working to reduce waiting times at ‘underperforming’ EDs has demonstrated that it is difficult to make a quick impact. Arguably the biggest impact we have had is changing the mindset of management about how to tackle the problem. An initial conversation about patient flow and ED waiting times held with a senior manager at one hospital illustrates this point. The following response was given when the author (TM) asked about in-patient processes further downstream. “Those [downstream processes] are out of scope. If we can just get the process right in the ED then we can move on”. By the end of the project the same individual recognized that the constraint for flow through the hospital was within the inpatient wards and outflow. This change in mindset and focus in improvement effort is important. In
Monks and Meskarian

the same hospital, we conducted a simulation project that lasted approximately three months. During that time the process in the ED changed three times at enormous emotional cost to the staff. This improvement effort was targeted in the wrong place.

8 WHERE NEXT FOR ED SIMULATION?

Two of the case studies presented had very short timescales. Results were needed in days as opposed to the typical simulation project timescale of months. Our experience with working hospitals in the South of England is that these timescales are common. As such, the time needed for detailed modelling (and associated data collection) to make an impact is rarely available. This is similar to the experience of Bowers et al. (2009) who found initial ‘simple’ analyses and work useful for an ED, but the time needed for more complex M&S meant that delivery was too late for impact.

The literature on the M&S of ED’s is extensive, as illustrated by a recent reviews (Gul and Gureni 2015; Saghafian et al. 2015). Within this literature much has been done to address the timeliness of simulation modelling of EDs by developing reusable generic models (e.g. Fletcher et al. 2007; Fletcher and Worthington 2009; Gunal and Pidd 2009). Our experience is that these models can rarely directly address the specific questions asked. However, such models may be relevant for the educational interventions illustrated by case study one. Fletcher et al. (2007) developed a ED model that they reused at 10 acute trusts in England. In many cases they could not get local data to parameterize the model or the process differed from their ‘generic’ model. However, they noted that there was benefit from simply getting the relevant parties together to discuss the general issues raised by the modelling related to their own system. We note that we have seen similar benefit from our educational interventions. There may be significant scope to consider a more educational approach for ED and hospital management using M&S.

A secondary issue preventing the literature on generic ED models making an impact is the need to recode them. To our knowledge none of the generic models published include a runnable version of the model. The most common type of simulation studies of EDs appear to be single site studies. Therefore the time needed to reproduce the models could be spent on producing a bespoke model more tailored to the study objectives. To improve uptake, it would appear then that M&S studies of EDs need to learn from Open Science (Baker 2016) and reduce the cost of reproducing reusable models. There are some encouraging signs that M&S is moving in that direction. For example, the journal ACM Transactions on Modelling and Computer Simulation now provides an optional reproducibility review for submitted models.

Lastly we again note the focus of M&S research on the ED, as opposed to the interconnected downstream hospital process (Saghafian et al. 2015). Although gains in ED performance can be made through micro-optimization of the ED process, we note that our three case studies highlight the need to consider ED outflow to the main hospital as the major constraint. It may be time ED M&S research to take a broader focus and analyze the relative effectiveness of solutions to increase the flexibility of downstream hospital processes.

9 CONCLUSIONS

Nearly all hospitals in the UK face a systemic problem that has led to bed shortages and ED overcrowding. The solutions to these systemic problems operate on a longer time cycle than the performance requirements on hospitals allow. As such, M&S offers a potentially high impact way to work with hospitals on increasing their flexibility and process efficiency. To do this M&S must meet two challenges. Firstly, identifying and testing opportunities to unblock ED admissions. Secondly, getting the level of detail and timeliness right to be useful for decision making in an extremely political and fast changing environment.
ACKNOWLEDGEMENTS

This article presents independent research funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) Wessex. The views expressed in this publication are those of the author(s) and not necessarily those of the National Health Service, the NIHR, or the Department of Health.

REFERENCES

Monks and Meskarian


AUTHOR BIOGRAPHIES

THOMAS MONKS is the Director of NIHR CLAHRC Wessex Data Science Hub. He is also a principal research fellow in the Faculty of Health Sciences, University of Southampton. He holds a BSc in Computer Science and Mathematics, MSc in Operational Research and PhD in Simulation Modelling. His speciality is simulation of healthcare systems of unscheduled and emergency care. His e-mail address is thomas.monks@soton.ac.uk.

RUDABEH MESKARIAN is a research fellow in Operational Research in the Faculty of Health Sciences, University of Southampton. She holds a BSc in Investment and Financial Risk Management, MSc in Operational Research and PhD in Operational Research concentrating on stochastic optimization. Her e-mail address is r.meskarian@soton.ac.uk.