A Queueing Network Model of Patient Flow in an Accident and Emergency Department

S.W.M. Au-Yeung, P.G. Harrison and W.J. Knottenbelt
Department of Computing
Imperial College London, SW7 2AZ, UK

August 15, 2006

ABSTRACT

In many complex processing systems with limited resources, fast response times are demanded, but are seldom delivered. This is an especially serious problem in healthcare systems providing critical patient care. In this paper, we develop a multiclass Markovian queueing network model of patient flow in the Accident and Emergency department of a major London hospital. Using real patient timing data to help parameterise the model, we solve for moments and probability density functions of patient response time using discrete event simulation. We experiment with different patient handling priority schemes and compare the resulting response time moments and densities with real data.

Introduction

It is a goal universally acknowledged that a healthcare system should treat its patients – and especially those in need of critical care – in a timely manner. However, this is often not achieved in practice, particularly in state-run public healthcare systems that suffer from high patient demand and limited resources.

In the United Kingdom, there has been much public concern regarding patient waiting times in the National Health Service (NHS). For example, in a recent King's Fund report, improved waiting times for patients in Accident and Emergency departments and for cancer and cardiac patients are identified as two of the public's top four priorities for public healthcare in the UK [9].

In response, the UK government has introduced performance targets for the NHS, many of which are driven by response times – in 2004/2005 NHS performance ratings were based on eight key targets, six of which involved patient waiting and treatment times. Currently NHS Trusts are assessed against a broader set of core standards, but these still incorporate existing response time targets. For example, 98% of patients should spend 4 hours or less in an Accident and Emergency department from arrival to admission, transfer or discharge. Although the vast majority of Acute trusts have managed to achieve a 95% threshold (assisted by innovations identified by the Emergency Services Collaborative such as "see and treat" schemes for minor injuries and near-patient testing [6]), 44% of Acute trusts are still failing to meet the 98% target [4]. This reflects the difficulty that many departments are experiencing in making further efficiency improvements [3]. This may be due, in part at least, to a lack of appropriate performance models and other systematic procedures for locating non-obvious capacity bottlenecks [6].

In this paper, we formulate a (simplified) hierarchical Markovian queueing network model of patient flow in the Accident and Emergency department of a major London hospital. Using real patient timing data to help parameterise our model, we compute moments and densities of patient treatment time using a discrete event simulation. We investigate the impact of giving priority treatment to different classes of patients, and compare the resulting response-time densities and moments with real data. We believe this work represents an important initial step towards the creation of a formal modelling environment for patient flow in hospitals that will allow hospital managers to assess the response-time impact of different resource allocations, patient treatment schemes and workload scenarios, and thereby to implement optimised patient flow pathways.

The idea of modelling health service departments is, of course, by no means new. Several studies have been made of patient flow in hospitals in general [7, 8, 15] and Emergency departments in particular [1, 2, 11, 12, 13, 5, 14]. However, these studies have had limited success and subsequent impact for two main reasons. Firstly, there has been a lack of sophistication in the models used (mostly simple discrete event simulations and very high-level queueing models), and in the analysis techniques applied (mostly aimed at computing resource measures such as utilisations and mean response times). Secondly, existing models frequently remain unvalidated using real waiting time data, since collecting this data was until recently a time-consuming, expensive, manual operation. We now have a prime opportunity to take advantage of the detailed patient waiting time data automatically collected by all Accident and Emergency (A&E) departments in England to monitor compliance with government targets (describing time of arrival, various treatment times and time of discharge

for every patient).
The remainder of this paper is organised as follows. The next section describes our multiclass Markovian queueing network model of patient flow. The numerical results section compares actual patient response times with our simulation results. The final section concludes and considers opportunities for future work.

Queueing network model

Description

Figs. 1 and 2 show the simplified multiclass queueing network model of patient flow we have developed in conjunction with an A&E consultant at our case study hospital. The model takes the form of a hierarchical network of $M/M/m$ queues. Fig. 1 shows top-level patient routing with various aggregated servers; their corresponding lower-level expansions are presented in Fig. 2. Our queueing model has four customer classes: patients with minor illnesses or trauma (minors), patients with major illnesses or trauma (majors), patients requiring resuscitation (resusc) and patients that have yet to be classified (assessment). Customers can change class as they proceed through the system. In the top-level model there are two forms of patient arrivals: walk-in patients who come into A&E via their own transport and patients that arrive by ambulance.

Walk-in Patients
These patients enter via the A&E waiting room where they are registered at reception. The receptionists route each patient into one of three queues: patients with a clear case of minor trauma are placed in the minors queue; patients with a clear case of a serious illness or serious trauma are sent to the majors queue; all others (including all suspected cases of minor illness), are sent for nurse assessment.

Minors Queue Patients in the minors queue must first wait for a minors cubicle to become free; the patient then waits there for a minors practitioner (either a minors doctor or a nurse practitioner) to see them. The minors practitioner can decide to:

- Perform investigative tests and/or scans such as blood tests and x-rays, or
- Ask for a specialist opinion, or
- Treat (if necessary) and discharge the patient (to home, their GP or to the pharmacy to pick up medication), or
- Send the patient to be admitted to a (surgical) ward, or the MAU (Medical Assessment Unit) which assesses the need for medical admissions.

Majors Queue Patients in the majors queue wait for a bed in a majors bay to become free; once there, a nurse may perform tests (e.g. vitals, blood tests, x-ray) so that essential information is ready for a doctor. When the doctor has assessed the patient, (s)he may require a specialist opinion, request more tests, or send the patient out of A&E (possibly after treatment) via the routes mentioned above for the minors queue. Occasionally a patient may suffer a sudden and rapid deterioration, in which case the patient is transferred to a resuscitation bay and is attended to by the resuscitation team. Tests for both majors and the minors are processed in the same laboratory and radiology facilities.

Nurse Assessment Patients in the nurse assessment queue wait for an assessment room to become available; they then wait there for a nurse who assesses the severity of their illness or injury. The nurse can send the patient either to the minors queue, the majors queue or discharge them out of A&E to a specialist clinic, ward, GP etc.

Specialists Specialists may be called in by a minors practitioner or majors doctor. Minors patients are only referred to "other" specialists which encompass ENT, Gynaecology and Orthopaedics. Majors patients may be seen by medical, surgical and "other" specialists. After assessment, patients are discharged from A&E, either being sent to a clinic for a more thorough investigation, being admitted to a ward or being sent to the MAU.

Ambulance Arrivals
These patients are handed over to a nurse from the ambulance. The nurse assesses the patient, decides which queue to assign them to, and sends them either to reception to be registered or straight to a majors bay (as appropriate).

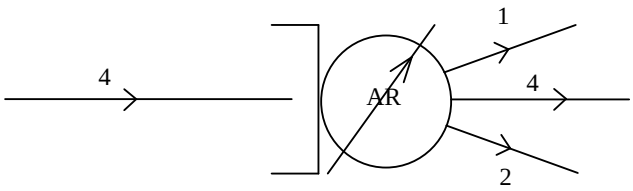
Blue Call Blue Call arrivals are very seriously ill or injured patients that require urgent medical attention. They almost always arrive by ambulance. Such patients are assigned a resuscitation bay and are attended to by a resuscitation team. Once stable the patient leaves A&E, being sent either to an operating theatre, to the ITU (Intensive Treatment Unit), or to a ward. Patients who cannot be resuscitated are sent to the mortuary.

Passive resources
Note that, in many cases a patient needs to obtain a (passive) resource before they can progress along a treatment path. An example is the nurse assessment rooms (of which there are 5 in our A&E department). A patient must wait for one to become free before entering the room for assessment by a nurse. Once the assessment has been completed, the patient leaves the room,

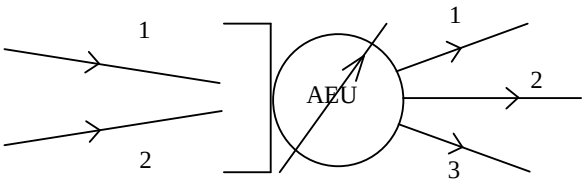
Patient Classes

1. Minors
2. Majors
3. Resusc
4. Assessment

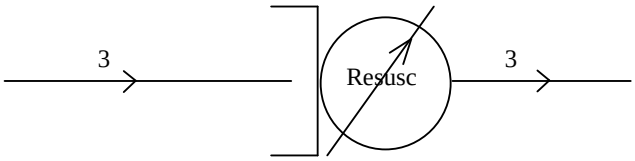
Aggregated Server AR (Assessment room)



Aggregated Server AEU (Whole medical unit)



Aggregated Server Resusc (Blue Call)



Top-Level Model

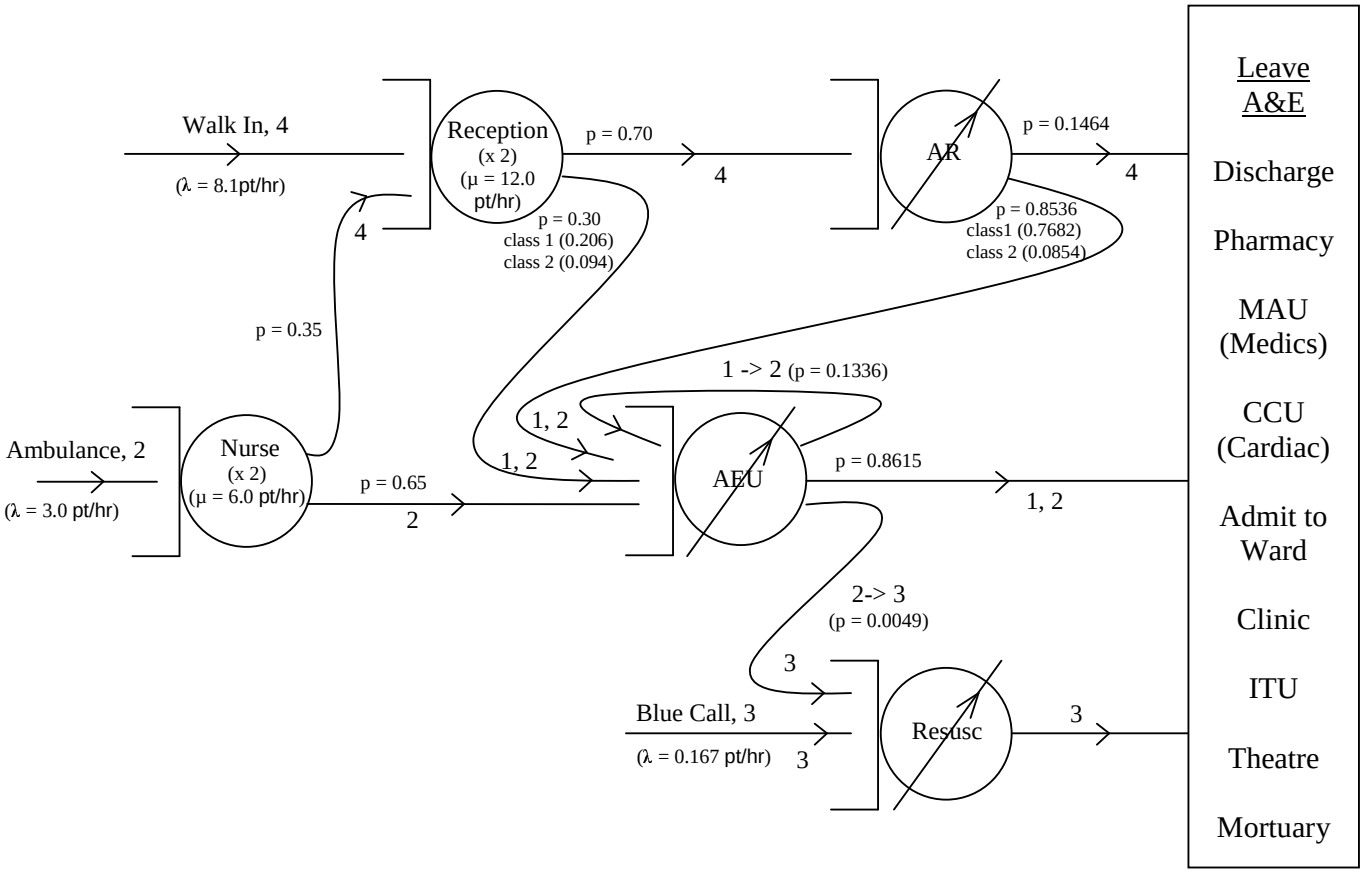
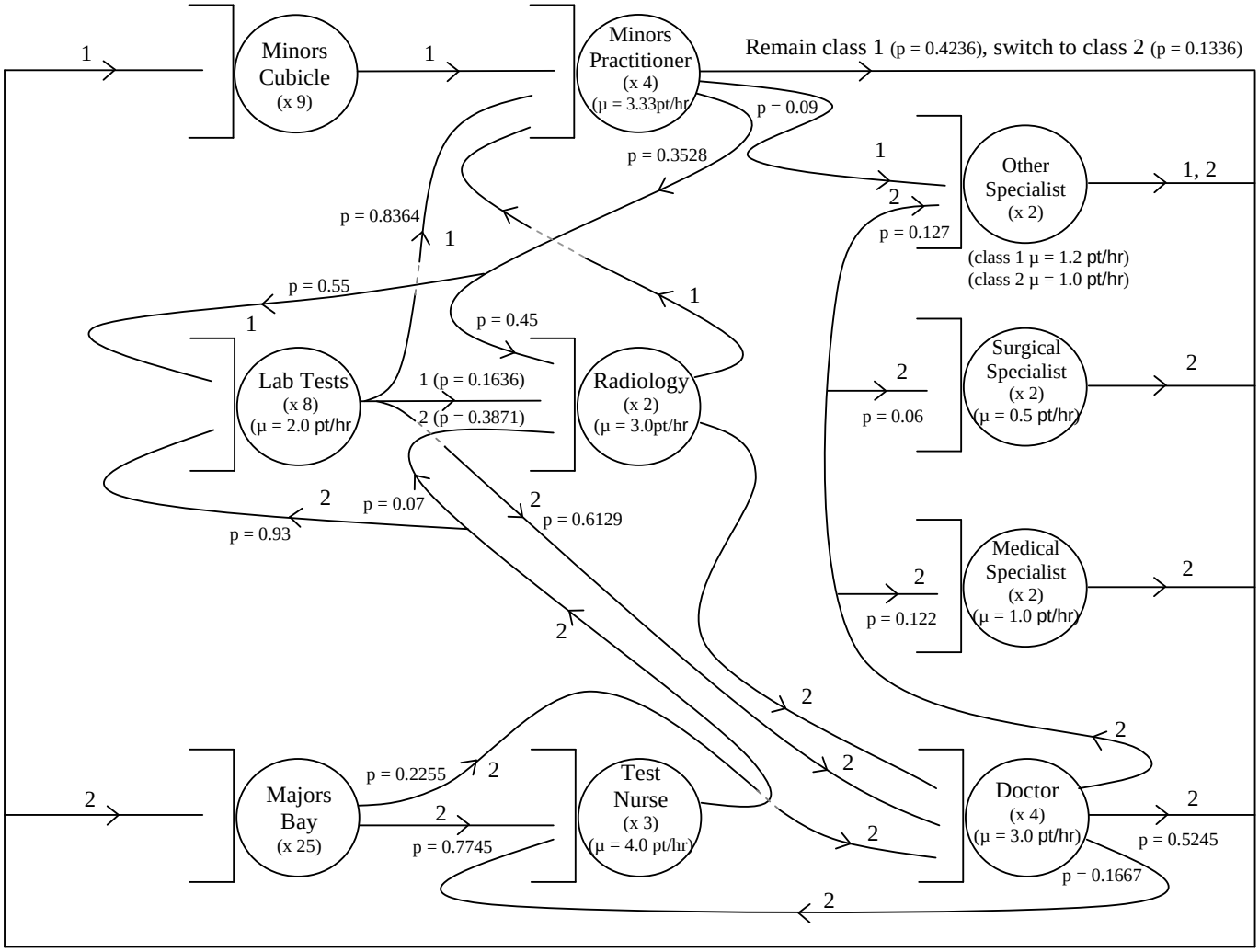
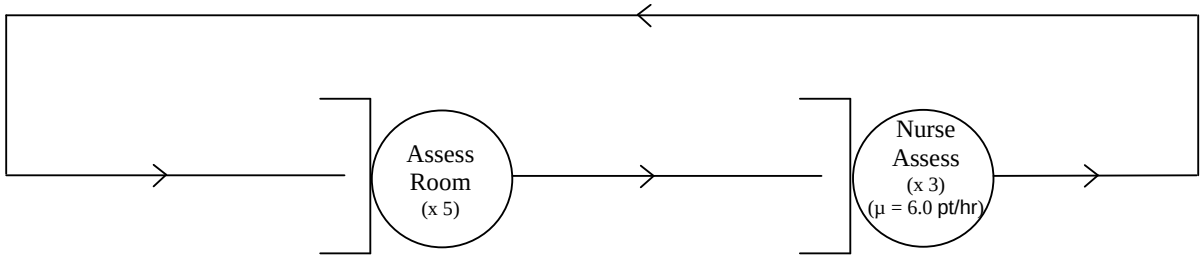


Figure 1: Top-level of queueing network model of patient flow

AEU Submodel



AR Submodel



Resusc Submodel

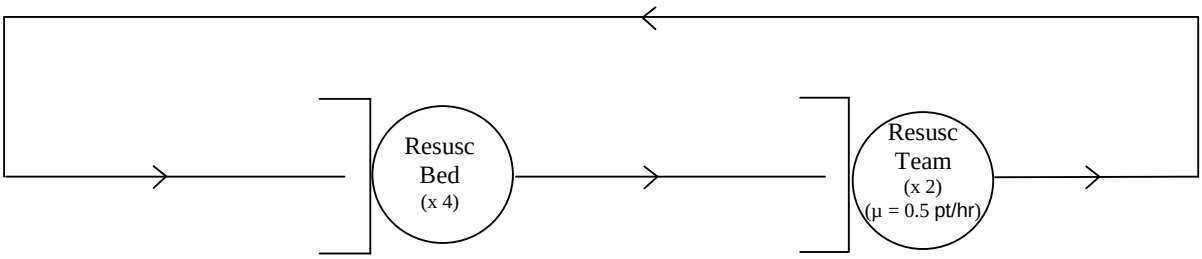


Figure 2: Lower-levels of queueing network model of patient flow

freeing it up for the next patient. Other passive resources include minors cubicles (of which there are 9), majors bays (of which there are 25) and resuscitation beds (of which there are 4).

Complexities not modelled
In a real-life A&E department there are many additional complexities that we have not incorporated into our model. For example, patients who should be discharged from A&E into another hospital ward are sometimes held in A&E even though their treatment in A&E is complete because there is no bed/room available for them in the destination ward. Similarly, patients may be held in A&E after completion of treatment awaiting an ambulance to take them home. We have not modelled these blocking phenomena caused by factors outside the A&E unit.
Patients who cannot walk must be transported around the A&E unit and taken to other areas of the hospital by porters; these are not represented in our model.
We have also had to simplify the nature of the tasks undertaken by various staff. For example, we have assigned nurses to perform specific tasks e.g. some nurses only assess patients. In a real A&E unit all the nurses are trained to perform assessments and treatments and so provide a more flexible staffing pool. As another example, there are many more types of specialist available in a real hospital than we have modelled. Also staffing levels and patient arrival rates vary throughout the day; we have used average values in order to simplify our model.
Finally, we have incorporated treatment time into the time it takes for a patient to be seen by either the doctor or minors practitioner. Depending on the nature of the patient's illness or injury, this may or may not be the case in an actual A&E unit.

Parameterising the model

We have obtained ethical approval to access detailed patient data timing collected by our case study Accident and Emergency department in a North London hospital. Where possible, we have used this data to parameterise our model. In particular, data for the year April 2004 to April 2005 was used to work out mean arrival rates: in that year there were 70 909 walk in arrivals and 26 285 ambulance arrivals; from experience there are 4 blue call arrivals a day – giving us mean arrivals rates of 8.1 walk in arrivals per hour, 3.0 ambulance (but not blue call) arrivals per hour and 0.167 blue call arrivals per hour. Where possible, we used the data to derive patient routing probabilities and mean service times; where this was not practical, we have used estimates provided by an A&E consultant, who has also checked the patient flows. Staff and resource numbers were provided by the hospital. Since there are different staffing levels throughout the day, we have taken average values (see Figs. 1 and 2 for staff numbers and service rates).

Numerical Results

We now compare numerical results from our discrete event simulation (written in Java) and real data.

Mean and variance of patient response time

Table 1 compares the first two (central) moments of patient response time for various types of patient arrival (Walk-in, Ambulance and Blue call arrivals) as calculated using our discrete-event simulation. The simulation results presented are the average of ten runs. Each run includes a transient period during which 2 000 000 patients move through the system (and during which passage time statistics are not collected), followed by a measurement period which lasts long enough to observe 10 000 passages of Blue Call arrivals through the system; in this period around 485 000 passages of Walk-in arrivals and 180 000 passages of Ambulance arrivals are also observed.
Three different patient priority schemes are analysed:

- *No Priority* in which First In First Out (FIFO) queues are implemented,
- *Majors Priority* in which majors patients are given priority at the shared resources (lab tests, radiology and "other" specialist), and
- *Minors Priority* in which minors patients are given priority at the shared resources.

From Table 1 it can be seen how giving priority to the majors class seriously degrades the waiting time of the walk-in patients (in terms of mean and variance), which are predominantly minors. By contrast it might appear that seriously injured or ill patients arriving by ambulance actually benefit from giving minors priority. In fact both ambulance and walk in arrivals under minors priority are seemingly treated quicker than even a no priority system. However, this interpretation may be misleading: a significant proportion of ambulance arrivals end up as minors (about 35%) and their benefit outweighs the penalty suffered by the majors that arrive by any means. Conversely, the walk-in major patients are highly penalised because relatively few walk-in minors patients switch to majors (about 16%). A separate comparison of ambulance arrivals that are treated as majors throughout their stay against walk-ins that are treated as minors throughout, i.e. neglecting any patients that change class, will reveal the true effects of changing between majors and minors priority. However, it must be remembered that the most important statistics to the individual patient concern their own time spent in hospital, regardless of the class to which they may be assigned.
Table 2 shows the first two moments of patient response time for various types of patient arrival (Walk-in, Ambulance and Blue call arrivals) as actually observed in

	Walk-In arrivals		Ambulance arrivals		Blue Call arrivals	
	E[T]	Var[T]	E[T]	Var[T]	E[T]	Var[T]
No Priority	2.86	8.57	2.77	5.28	2.08	4.19
Majors Priority	5.15	37.22	3.48	17.19	2.06	4.12
Minors Priority	2.05	4.05	2.63	4.82	2.07	4.15

Table 1: Mean and variance of response times (in hours) for walk in, ambulance and blue call patients under major, minor and no priority schemes.

	Walk-In arrivals		Ambulance arrivals		Blue Call arrivals	
	E[T]	Var[T]	E[T]	Var[T]	E[T]	Var[T]
2002/2003	3.22	13.03	5.69	23.40	4.18	26.95
2003/2004	2.46	4.98	4.22	9.73	2.43	4.81
2004/2005	2.04	2.54	3.14	4.49	2.09	3.37

Table 2: Observed mean and variance of response times (in hours) for different classes of arriving patient.

the A&E we have modelled. Figures are reported over three annual reporting periods (2002/2003, 2003/2004 and 2004/2005), where each reporting period begins on 1 April and ends on 31 March the following year (coinciding with the hospital's financial year). One can readily observe the effect of the introduction and subsequent tightening of patient response time targets. The practical effect of this has been to move from a system in which majors are given priority treatment to a system in which minors are (to a large degree - since the majority of patients in A&E departments are minors patients) given priority treatment. Indeed, the trends observed (with associated reductions in the mean and variance of patient waiting time) are consistent with those we observe when moving from a majors priority to a minors priority schemes (cf. Table 1).

When comparing mean patient response times from our minors priority simulation model with the observed 2004/2005 figures, we observe differences of 0.5%, 16.2% and 1% for Walk-in, Ambulance and Blue call patients respectively. The relatively large disagreement between ambulance arrival actual treatment time and our simulation may be due to the lack of blocking phenomena in our model, which will mostly delay ambulance arrivals. However, the close agreement for Walk-in and Blue call arrivals is promising, considering the many simplifying assumptions we have made.

Densities of patient response time

Figures 3, 4 and 5 present plots of simulated vs. actual patient response time densities for Walk-in, Ambulance and Blue call arrivals respectively; note that the curves corresponding to the no priority system lies in between the curves for the majors and minors priority systems, also note the peaks in the 2004/2005 actual patient response time densities which correspond to the four hour target.

Conclusion

In this paper, we have used a simulation modelling to provide some insights into the effects of prioritising different classes of patients in a real Accident and Emergency unit based in London, UK. We have found that the (seemingly socially unacceptable) prioritisation of treatment for minors (i.e. patients with minor illness or trauma) over majors (i.e. patients with severe illness or trauma) can lead to the counter-intuitive outcome that mean response times for ambulance arrivals are not adversely affected (in fact they are slightly improved), while mean response times (and corresponding variances) for walk-in arrivals are dramatically lower. This is a particularly interesting result in light of UK government waiting time targets, which encourage the prioritisation of minors.

In the future, we intend to incorporate more realistic assumptions into our models. For example, the arrivals process at a real hospital is non-stationary and is more bursty than a Poisson arrivals stream. Also, our model does not yet represent the "rising panic" phenomenon that occurs in real A&E units whereby patients are subject to higher and higher priority treatment as they approach the four hour waiting time target. Some progress towards modelling queues where the priority of a customer increases with the time spent in-queue was made by one of the present authors in [10]. There a queue was represented by an ordered set of current customer *sojourn times*. This has led to a uniform way of deriving response time distributions under various queueing disciplines and a rather complex, untested approximate route to deadline queues. We will tailor this approximation to our problems.

Acknowledgements

We are grateful for the help and advice of many members of staff at our case study hospital and associated

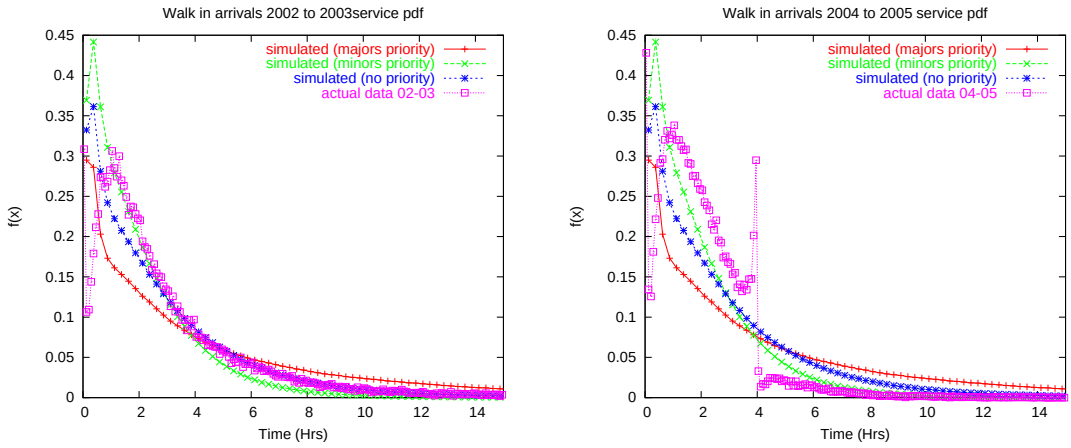


Figure 3: Actual and simulated response time density for walk-in arrivals using 2002/2003 data (left) and 2004/2005 data (right)

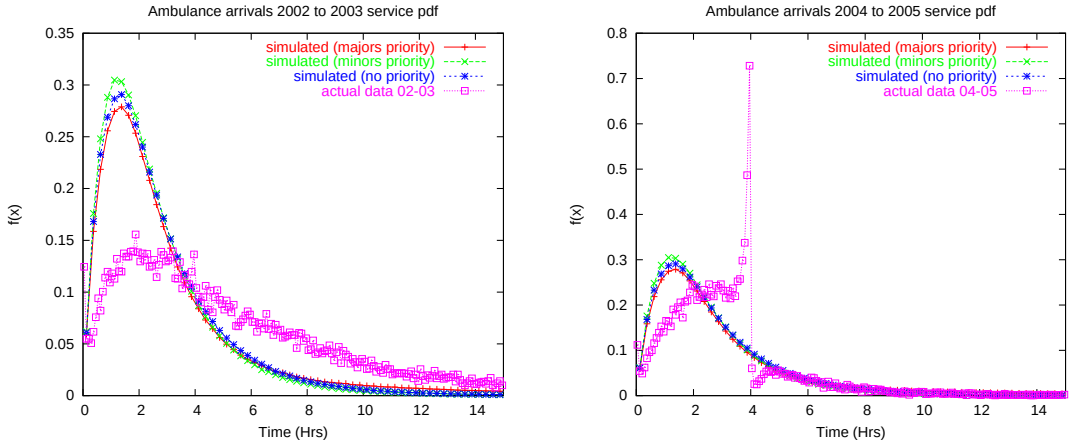


Figure 4: Actual and simulated response time density for ambulance arrivals using 2002/2003 data (left) and 2004/2005 data (right)

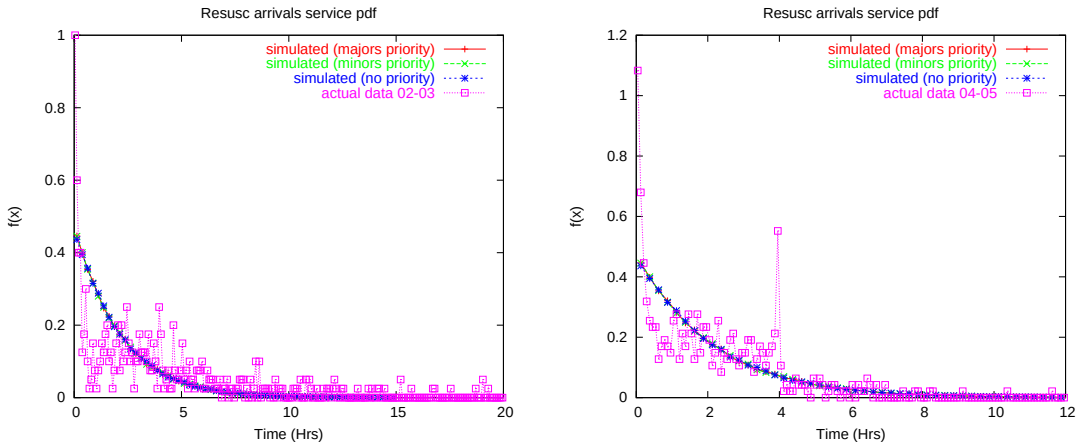


Figure 5: Actual and simulated response time density for blue call arrivals using 2002/2003 data (left) and 2004/2005 data (right)

institutions, including John Knottenbelt, Rick Juniper, David King, Raj Singh, Sunil Johal, Sharon Ahearn and Ken Walton. We would also like to thank Tony Field for the use of his JINQS Java queueing network simulation library. Ethical approval for access to psuedonymised patient records was granted by the Harrow Local Research Ethics Committee (Ref. 04/Q0405/72).

REFERENCES

[1] J.T. Blake and M.W. Carter. An analysis of Emergency Room wait time issues via computer simulation. *Information Systems and Operational Research (INFOR)*, 34(4):263–273, November 1996.

[2] T.J. Coats and S. Michalis. Mathematical modelling of patient flow through an Accident and Emergency department. *Emergency Medicine Journal*, 18:190–192, 2001.

[3] Healthcare Commission. Acute hospital portfolio review. accident and emergency. Technical report, August 2005.

[4] Healthcare Commission. NHS performance ratings 2004/05. Technical report, 2005.

[5] L.G. Connelly and A.E. Bair. Discrete event simulation of emergency department activity: A platform for system-level operations research. *Academic Emergency Medicine*, 11(11):1177–1185, 2004.

[6] M. Cooke, J. Fisher, and J. Dale et al. Reducing attendances and waits in emergency departments a systematic review of present innovations. Technical report, Report to the National Co-ordinating Centre for NHS Service Delivery and Organisation R & D (NCCSDO), 2005.

[7] Murray J. Côté and William E. Stein. An Erlang-based stochastic model for patient flow. *Omega: The International Journal of Management Science*, 28:347–359, 2000.

[8] R. Davies and H.T.O. Davies. Modelling patient flows and resource provision in health systems. *Omega: The International Journal of Management Science*, 22:123–131, 1994.

[9] The King's Fund. Has the government met the public's priorities for the NHS?: A King's Fund briefing for the BBC 'Your NHS' day 2004. Technical report, 2004.

[10] P.G. Harrison. An M/M/1 queue with aging priority. In *Proc. International Conference on Stochastic Modelling and the IV International Workshop on Retrial Queues*, Cochin, India, December 2002.

[11] D. Lane, C. Monefeldt, and J. Rosenhead. Emergency – but no accident – a systems dynamics study of an Accident and Emergency department. *OR Insight*, 11:2–10, 1998.

[12] L. Mayhew and E. Carney-Jones. Evaluating a new approach for improving care in an Accident and Emergency department: The NU-care project. Technical report, Cass Business School, City University, 2003.

[13] Ò. Miró, M. Sánchez, G. Espinosa, B. Coll-Vinent, E. Bragulat, and J. Millá. Analysis of patient flow in the emergency department and the effect of an extensive reorganisation. *Emergency Medical Journal*, 20:143–148, 2003.

[14] A. C. Virtue. Simulating accident and emergency services with a generic process model. Nosokinetic News, December 2005.

[15] E.N. Weiss, M.A. Cohen, and J.C. Hershey. An iterative estimation and validation procedure for specification of semi-Markov models with application to hospital patient flow. *Operations Research*, 30(6):1082–1104, 1982.