Server Behaviours in Healthcare Queueing Systems

Paul. R. Harper
School of Mathematics, Cardiff University, Cardiff, Wales, UK.
harper@cardiff.ac.uk

Server Behaviours in Healthcare Queueing Systems

Abstract
In the classical queueing theory literature, a server is commonly assumed to work at a constant speed. Motivated by observations from healthcare applications, a study is made to explore the nature of the relationship between service times and workload in order to assess and quantify any workforce (server) behaviours. Consequently, an initial analytical queueing model is considered with switching thresholds to allow for two-speed service. In this model service time depends on queue length, which for example captures the congestion in the waiting room and the resulting change in speed of the workforce to try and cope with the backlog of patients. Furthermore, related behavioural characteristics resulting from workload fatigue and service breakdown are considered. A developed analytical model with 'catastrophic' service failure is proposed to examine the consequences on patient service levels. The research helps to demonstrate the importance of more accurately capturing server behaviours in workload-dependent environments, and the impact this has on the overall system performance. It is hoped that this paper might help spawn an emergence of behavioural queueing systems literature, retaining the role that queueing theory plays within our field but with enhanced consideration of the role of behaviours when constructing such models.

Keywords
Behavioural OR, queueing systems, healthcare, workload-dependent service.

1. Introduction

There has been an emergence of interest in Behavioural Operational Research (BOR) in recent years (Hämäläinen et al 2013). BOR studies are designed to advance our understanding of how behavioural factors affect the conduct of, and interaction with, model-based processes that support problem solving and decision making (Franco and Hämäläinen, 2016). Whilst the majority of recent interest in BOR has tended to focus on client-engagement and stakeholder behaviours throughout the modelling process, as well as the development of agent-based simulations for reflecting behaviours within OR models, the contribution of this paper is to consider how we might better capture human behaviours within queueing theory based models.

In the classical queueing theory literature, a server is commonly assumed to work at a constant speed. That is, we implicitly assume the server's productivity is independent of the workload. However there are many real-world observations when such an assumption may not be appropriate, such as varying service speed in production lines based on orders placed and product demand. Kc and Terwiesch (2009) study the impact of workload on service time and patient safety within hospital operations, specifically for cardiothoracic patients. They find that workers (healthcare staff) accelerate the service rate as load

increases. In particular, a 10% increase in load reduces length of stay by 2 days. Jaeker and Tucker (2016) study two years of inpatient data from 203 Californian hospitals and observe that patient length of stay increases as occupancy increases, until a tipping point, after which patients are discharged early to alleviate congestion. Moreover, the authors find a second tipping point beyond which additional occupancy leads to a longer length of stay, indicative of a workload-related "saturation effect" where employees can no longer overcome high workload by speeding up. Other studies have reported workload-dependent service across different customer intensive environments such as Anand *et al* (2011), Tan (2014) in restaurants, and further work by Kc and Terwiesch (2013) in healthcare.

Motivated by findings from an empirical study of a large emergency department, this paper builds on the literature to more formally consider models for explicit consideration of situations when the time it takes a resource to serve a patient depends on the current state of that queueing system, specifically the workload as measured by the current queue length for service. There is a vast and growing literature on the use of OR in healthcare service operations, including comprehensive reviews such as those by Brailsford *et al* (2009) and Hulshof *et al* (2012). It is fair though to say that the majority of studies that capture the true underlying stochastic nature of such systems (Harper, 2002) traditionally fit distributions to observed service times, such as length of stay, and use these within developed analytical, or more typically, simulation models. Hence the majority of work to-date implicitly fails to consider any relationships between workforce behaviours/productivity and workload. Of course, it may well be the case that such a relationship does not pertain to the healthcare system under investigation but since this is rarely, if ever, reported, we may assume that it has not been considered. Therefore it is currently not known if such an assumption is detrimental to the quality of results and decision-making for resourcing levels.

In this paper an initial analytical queueing model is consider for switching thresholds to allow for two-speed service that better reflects workforce productivity. Furthermore, consideration is given to the related issue of staff burnout caused by sustained levels of high workload and the resulting impact on patient service. The overarching aim of this research is thus to explore whether more detailed modelling of server behaviours in queueing systems, and the necessary effort to do so, is worthwhile in providing greater precision in capacity planning decision making compared to the use of more commonly used methods that assume service rates to be exogenous of resource utilisation.

Whilst the models presented in this paper may still be considered as a simplified version of reality and not fully able to capture all aspects of exhibited behaviours, they are intended to provide helpful insights, tools to quickly compute key performance metrics, and to incorporate the major features evidenced in the motivating case study. The contribution of this work is thus not tied to the proposed models being an entirely accurate description of how a server responds to queue size, but rather the value is in having analytical models to explore the potential impact of such features of queueing in healthcare systems, and to hopefully motivate further research in this area.

The remainder of this paper is organised as follows. In section 2 we present findings from an empirical study of service times for a large emergency department. In section 3 an M/G/1-type queueing model is formulated to accommodate the observed switching threshold for two speeds of service. Results from the analytical insights and a developed simulation model are reported and compared in section 4. Section 5 considers workload fatigue and proposes an analytical model to capture service breakdown. We conclude with discussions and possible future research directions in section 6.

2. Motivating Empirical Study

Many healthcare systems across the globe are facing a time of austerity, having to deal with increasing demand and complexity in heath needs within constrained budgets. Designing and delivering prudent healthcare services to ensure resources are used to maximum effect is a challenging yet vital task. Improved understanding of patient and staff behaviours is therefore critical.

The use of techniques such as agent-based simulation can help facilitate this, and there has been an emergence of papers in this field. A good overview of applications to healthcare can been seen in Barnes *et* al (2013). However such methods still require rules to be assigned to the individual autonomous agents (for example both patients and healthcare staff) that govern how they react to changing environments, such as patient choice (Knight and Harper, 2013). To-date though, few healthcare OR papers have explicitly considered workforce behaviours. One can find other examples on more general topics such as organisational structures (e.g. Fetta *et al*, 2012).

The first known empirical study to demonstrate how healthcare employees adjust their service rate with changing levels of load was by Kc and Terwiesch (2009). They used data from a cardiothoracic surgery unit in a major US teaching hospital. The authors observed a clear pattern indicating that length of stay decreased with an increase in workload. By staff working faster, the unit increases its throughput when it is busy. The implications of their study is that the adaptive behaviour of the healthcare workforce increases the overall process flow of patients from the hospital that one might have not have otherwise observed if assuming a fixed single service time distribution.

Motivated by Kc and Terwiesch (2009) and Jaeker and Tucker (2016), and through discussions with collaborating clinical staff in different healthcare settings, patient data has been acquired from a large emergency department in Wales, UK. The data covers a period of 6 months (July 2015 – December 2015). Each patient record used in the analysis provides their time of arrival into the department, service time (taken as the time treatment commences to the time a decision was made, which could be discharge or a transfer to assessment unit or ward), and their triage category (an indicator of urgency/medical need). Hence it is possible to explore service times against patient census counts, defined here

as the number of patients waiting in the emergency department for service at the time the patient starts service.

Service time in the emergency department will typically be affected by medical need (case-mix) and possibly on congestion in other parts of the hospital, such as timely access to necessary scans, laboratory tests etc. In turn access to such resources itself may be time-dependent, such as limited staff for some additional tests during the night shift or over weekends. We therefore focus our empirical study on one shift (8am – 4pm) only, which has the advantage of comparing service times in daytime hours with a similar number of staff rostered on duty each day and when access to other related resources is typically less variable. Furthermore we consider only service times for those patients assigned in the *urgent* category of care. This has the advantage of best reflecting what the true workforce service time might be and removes the issue of having to adjust for risk/severity. Furthermore this category has the most number of attendances, the other categories being *critical* (typically taken immediately into care without any wait) and *non-urgent*. Whilst we acknowledge this is still by no means perfect, and even if in reality the recorded service times may still be influenced by exogenous factors, we do have a sufficiently large number of patients in the analysis (n = 4,832) to allow a relative like-for-like comparison of service times and congestion levels.

A violin plot (showing the shape of the distribution alongside the more traditional box plot format) of patient service times for each observed number of patients waiting at that time for commencement of service (patient census) is shown in Figure 1. The solid line connects the average service times across the range of patients waiting; note that in fact the highest observed number of patients waiting for service in the dataset was 24, but given the small counts above 13, the plot shows only up until this number. Figure 2 plots just the average service time and clearly shows the trend over the number of patients waiting.

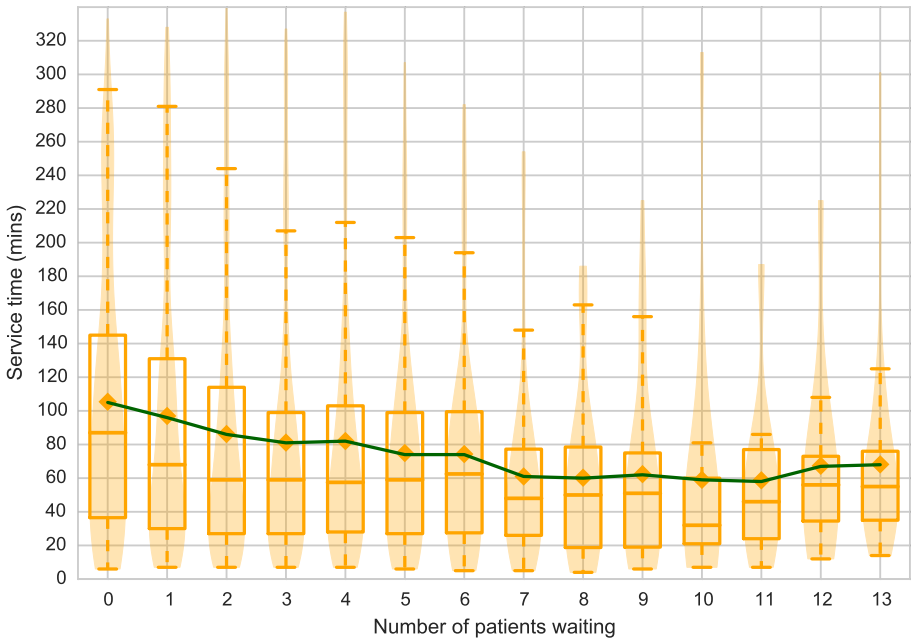

Figure 1: Service times as a function of number of patients waiting for service

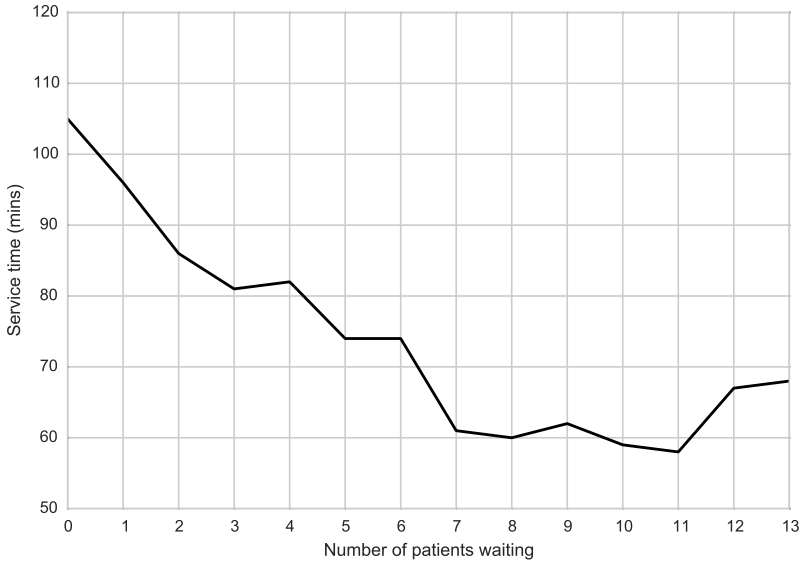

Figure 2: Average service time against number waiting for service

Figure 1 and 2 are very interesting in several respects. Generally it reinforces the findings of published work that service times change with a changing workload. However, rather than a smooth continuous relationship over the range of workload, there appears to be a possible threshold at which we observe a levelling off in service rates (around 7 patients waiting for service in the emergency department). Furthermore, for the highest levels of workload (in excess of 11 patients waiting), there is an apparent increase again in service

times, or conversely a drop-off in staff productivity. Interestingly, this is similar to reported studies that have also demonstrated hospital resources cannot sustain the increased service rate and that there may exist an effect of overwork (Kc and Terwiesch, 2009), or what Jaeker and Tucker (2016) term a workload-related "saturation effect" where staff can no longer overcome high workload by speeding up. Indeed medical studies also report of increased productivity in the short run followed by burnout after a prolonged period of exceptionally high service rate activity; see for example Aiken *et al* (2002), Gaba and Howard (2002) and Rubulotta *et al* (2016).

Finally, although not frequently observed and therefore inconclusive, there also appears to be the possibility of longer service times at very low levels of activity (0 or 1 patients waiting). It would therefore seem that when the unit is particularly quiet, staff are not hurried and thus slow down their service speed. Whilst we cannot provide a definitive explanation, it could be explained by staff slowing down to fill the time they have available, or staff providing the necessary levels of patient care that perhaps aren't otherwise fully given when rushed.

In summary, Figure 1 would suggest that for the emergency department under investigation:

- As the unit moves from low workload levels to more moderate levels, service times decrease;
- As the unit moves from moderate to high workload levels, there appears to be a possible step-change with service times decreasing and then levelling off, possibly indicating a switch in productivity to handle the increasing backlog of patients;
- As the unit moves from high to exceptionally high workload levels, service times increase again, possibly indicating workforce burnout, low morale give the levels of congestion and busyness, and/or physical space constraints limiting the movement of patients, staff, and resources.

Given that EDs largely operate as a pre-emptive priority queueing system, further analysis of the data explored any influencing factors caused by the higher acuity *critical* patients, given medical staff will generally leave an urgent case to treat a critical one. According to the acuity classification adopted by the hospital, there were only 172 critical admissions (compared to the 4,832 urgent cases) and there were no unusual peaks in critical demand that might have particularly influenced service times for urgent patients. Hence we conclude that with the relatively low critical demand evenly distributed over the time period under study, that the relationship between service times and workload is seemingly not attributable to pre-emptive factors relating to higher acuity demand.

In subsequent discussions with ED staff, they sensed that both short-term and longer–term burnout issues were potential influencing factors. In the short-term (such as a single shift), at times with particularly high demand resulting in higher numbers of patients waiting, a shift may be prolonged (working overtime, either officially or unofficially; some staff may decide to stay on beyond their end of shift to help colleagues cope with the backlog). With successive periods of high workload (such as over weeks), a repeated pattern of busy and long shifts may

cause longer-term burnout and low morale, which can consequently contribute to staff illness and absence rates. In turn this can result in financial implications for having to bring in more expensive temporary staff to cover absence. To better understand the reasons for the observed relationship (Figures 1 and 2), clearly this study would benefit from further inter-disciplinary research including qualitative methods and insights drawn from ethnography and the social sciences.

3. A Workload-Dependent Queueing Model

When reviewing the literature on workload-dependent service rates, the pioneering work in this field is by Satty (1961) and Gebhard (1967) who consider M/M/1 queues. Subsequent work has considered multi-server systems, with Garg and Singh (1993) determining the optimal queue length at which to employ a second server in an M/M/2 system. Wang and Tai (2000) extend this for M/M/3 and Lin and Ke (2011) use genetic algorithms to find the best thresholds for change in number of servers for M/M/c systems.

More recent research has considered modelling queueing systems with adaptable service rates, such as by Zhernovyi (2012) who examines the stationary characteristics of a $M^X/M/1$ system with two-speed service, and by Baër *et al* (2014) for a *PH*/*PH*/1 multi-threshold model. Tirdad *et al*. (2016) consider optimal control points of M(t)/M/c/c queues with periodic arrival rates and two levels of the number of servers and apply their model to an emergency room at the Kelowna General Hospital, US.
In this paper we initially consider an M/G/1 system, motivated by the fact that it is well documented that many service times (lengths of stay) in healthcare processes are rarely exponentially distributed (see for example Faddy *et al*, 2009) and indeed as observed in the ED case study at hand. Hence for flexibility, it is preferable to accommodate more general length of stay distributions.

Motivated by the empirical study above and of similar previously published findings, an analytical queueing model, namely an M/G/1-type model is now considered, which is able to capture some of the observed server behaviours as opposed to classical queueing theory models with assumed constant service rates. The data and insights from section 2 certainly provide further evidence of a major feature (lower service times at higher queue lengths) absent from traditional queueing models. Whilst the model presented here may still be considered as a simplified version of and not fully able to capture all aspects of Figure 1 behaviours, it is intended to incorporate this major feature, provide helpful insights, and to quickly compute key performance metrics.

Initially, we assume a single server and two-speeds of service, with thresholds (numbers of patients waiting for service) at which the service rate switches. Figure 1 would seem to indicate more than 2 potential speeds, although for initial insights we construct a two-speed model given already the well documented challenges of deriving the necessary equations for any more than this, particularly given here we are incorporating non-Markovian service time distributions.

Building on Gray and Wang (1992), here we incorporate two thresholds: an adaptive server will change speed when the number of patients in the emergency department reaches B and will retain this speed until such a time when the number of patients waiting reduces below A. For a single switching threshold, A and B would be set to the same value, but the formulation here allows for the flexibility for the higher server rate to be in operation longer. This may be particularly useful in healthcare settings where staff continue to work faster until the backlog of patients is cleared and the number still waiting for service is below the original threshold. Once threshold B is again reached, the service speed is again switched, and so on. Furthermore in this paper we consider both mean number and waiting time of patients in the queue for service.

Patients (or more broadly customers, since this work could equally be applied to different service settings) arrive according to a Poisson distribution with rate λ. Arrival and service times are independent of each other. Patients are served on a FCFS basis. When there are no more than B patients waiting in the queue ($B >$ 1), service times are i.i.d random variables with associated density $f_1(x)$ for service speed 1. If upon completion of a service there are more than B patients in the queue, then the service time switches to follow a different distribution $f_2(x)$ for service speed 2. Service speed 2 remains in effect until if on completion of a service the queue length is reduced to A ($0 \leq A \leq B$), then the service speed switches back again to service speed 1.

Let S_i be a random variable corresponding to $f_i(x)$ i = 1, 2. S_i has a mean of $1/\mu_i$ and variance σ_i^2. Let $\rho_i = \lambda/\mu_i$ and that $\rho_i < 1$. Let U_i represent the number of patients who arrive during a service time S_i.

Let

$$u_n = P(U_1 = n) = \int_0^\infty f_1(t) \frac{(\lambda t)^n}{n!} e^{-\lambda t} dt \qquad n = 0, 1, \ldots$$

and

$$v_n = P(U_2 = n) = \int_0^\infty f_2(t) \frac{(\lambda t)^n}{n!} e^{-\lambda t} dt \qquad n = 0, 1, \ldots$$

Let $U(z)$ and $V(z)$ be the generating functions of $\{u_n\}$ and $\{v_n\}$ respectively. Then:

$U'(1) = \rho_1 \qquad U''(1) = \lambda^2\sigma_1^2 + \rho_1^2$

$V'(1) = \rho_2 \qquad V''(1) = \lambda^2\sigma_2^2 + \rho_2^2$

We analyse this model as an embedded Markov chain, in which the states of the chain may be divided into two classes as follows:

$V_1 = \{(n, 1); n = 0, \ldots, B\}$

where $(n, 1)$ represents the state in which n patients are still in the queue at the end of a service and the next service is to be of service speed 1, and

$V_2 = \{(n, 2); n = A + 1, A + 2, \dots\}$

where $(n, 2)$ represents the state in which n patients are still in the queue at the end of a service and the next service is to be of service speed 2.

Let us denote the stationary probability of the state (n, i) by $\pi(n, i)$. The empty state of the system thus corresponds to (0,1) and is denoted by π_0. We now derive the equations for the stationary probabilities of the chain as follows:

$$\pi(i,1) = u_i\pi_0 + \sum_{j=1}^{i+1} u_{i-j+1}\pi(j,1) \qquad 0 \le i \le B-1, \qquad i \ne A \tag{1}$$

$$\pi(A,1) = u_A\pi_0 + \sum_{j=1}^{A+1} u_{A-j+1}\pi(j,1) + v_0\pi(A+1,2) \tag{2}$$

$$\pi(B,1) = u_B\pi_0 + \sum_{j=1}^{B} u_{B-j+1}\pi(j,1) + v_0\pi(B+1,2) \tag{3}$$

$$\pi(A+i,2) = \sum_{j=1}^{i+1} v_{i-j+1}\pi(A+j,2) \qquad i = 1, \dots, B-A \tag{4}$$

$$\begin{aligned}\pi(B+i,2) = {} & u_{B+i}\pi_0 \\ & + \sum_{j=1}^{B} u_{B+i-j+1}\pi(j,1) \\ & + \sum_{j=1}^{B-A+i+1} v_{B-A+i-j+1}\pi(A+j,2) \qquad i = 1,2,\dots\end{aligned} \tag{5}$$

Define the following generating functions:

$$\pi_1(z) = \sum_{i=0}^{B} \pi(i,1)z^i$$

and

$$\pi_2(z) = \sum_{i=A+1}^{\infty} \pi(i,2)z^i$$

so that from equations (1) – (5) we obtain the following relationship:

$$(U(z) - z)\pi_1(z) + (V(z) - z)\pi_1(z) = U(z)(1-z)\pi_0 \tag{6}$$

and we require that:

$$\pi(1) + \pi(2) = 1 \tag{7}$$

Equations (6) and (7) alone are insufficient to determine π_0 and the mean queue length, and hence it is required therefore to determine $\pi_1(1)$ and $\pi_1'(1)$. This is achieved through the following algorithm.

Let $\pi(n,1) = \pi_n \pi_0,\ 1 \le n \le B$. Let $\pi(n,2) = \varphi_n \pi_0,\ n \ge A+1$. Also let π_n' $(0 \le n \le B)$ be the coefficient of the probability of the empty state in a regular M/G/1 queue with service time density $f_1(x)$. We note that the π_n' satisfy:

$$\pi_0' = 1 \qquad \pi_1' = (1-u_0)/u_0 \tag{8}$$

and

$$\pi_{i+1}' = \frac{\left[(1-u_1)\pi_i' - u_i\,\pi_0' - \sum_{j=1}^{i} u_{i-j+1}\,\pi_j'\right]}{u_0} \qquad i = 1, \dots, B-1 \tag{9}$$

We can write π_n in the form:

$$\pi_n = \pi_n' \qquad 0 \le n \le A \tag{10}$$

and

$$\pi_{k+1} = \pi_{k+1}' + \omega_i \varphi_{A+1} \qquad i = 1, \dots, B-A \tag{11}$$

where ω_i $(i = 1, \dots, B-A)$ are computed from equations (1) – (5). From these equations we note that ω_i satisfy the following recursive relations:

$$\omega_1 = -\frac{v_0}{u_0} \tag{12}$$

and

$$\omega_i = \frac{\left[(1-u_1)\omega_{i-1} - \sum_{j=1}^{i-1} u_{i-j}\omega_j\right]}{u_0} \qquad i = 2, \dots, B-A \tag{13}$$

We can compute the quantity φ_{A+1} by:

$$\varphi_{A+1} = \frac{\left[(1-u_1)\pi_B' - u_B\pi_0' - \sum_{j=1}^{B-1} u_{B-j+1}\pi_j'\right]}{D_{B,A}} \tag{14}$$

where

$$D_{B,A} = \sum_{j=1}^{B-A-1} u_{B-A-j+1}\omega_j - (1-u_1)\omega_{B-A} \qquad A < B \tag{15}$$

Note that in the special case of $A = B, D_{B,A} = v_o$

We now set $\frac{\pi_i(z)}{\pi_0} = \widehat{\pi_\iota}(z), i = 1, 2.$ From equations (6) – (8), it follows that:

$$(U(z) - z)\widehat{\pi_\iota}(z) + (V(z) - z)\widehat{\pi_2}(z) = U(z)(1 - z) \tag{16}$$

and

$$\pi_0 = 1/[\widehat{\pi_\iota}(1) + \widehat{\pi_2}(1)] \tag{17}$$

Using equations (8) – (15) we have:

$$S = \widehat{\pi_\iota}(1) = \sum_{i=0}^{B} \pi_i' + \left[\sum_{i=1}^{B-A} \omega_i\right] \varphi_{A+1} \tag{18}$$

and

$$T = \widehat{\pi_\iota'}(1) = \sum_{i=0}^{B} i\pi_i' + \left[\sum_{i=1}^{B-A} (A + i)\omega_i\right] \varphi_{A+1} \tag{19}$$

By differentiating (16) and setting z = 1, an expression for $\widehat{\pi_2'}(1)$ is found. Then π_0 is found by using (17):

$$\pi_0 = \frac{1 - \rho_2}{1 + S(\rho_1 - \rho_2)} \tag{20}$$

Finally, the mean queue length is found by differentiating (6) twice, setting z = 1 and using equations (18) – (20). Hence we derive the expected queue length and waiting time for a system with a two-speed server, as:

$$L_q = \frac{\lambda^2\sigma_1^2 + \rho_2^2}{2(1 - \rho_2)} + \frac{2\rho_1 + S[\lambda^2(\sigma_1^2 - \sigma_2^2) + \rho_1^2 - \rho_2^2] + 2T(\rho_1 - \rho_2)}{2[1 + S(\rho_1 - \rho_2)]} \tag{21}$$

$$W_q = \left(\frac{\lambda^2\sigma_1^2 + \rho_2^2}{2(1 - \rho_2)} + \frac{2\rho_1 + S[\lambda^2(\sigma_1^2 - \sigma_2^2) + \rho_1^2 - \rho_2^2] + 2T(\rho_1 - \rho_2)}{2[1 + S(\rho_1 - \rho_2)]}\right) / \lambda \tag{22}$$

For a comparison of the mean queue length and waiting time from our adaptive behaviour server, we may compare L_q and W_q to standard results for a M/G/1 queueing system:

$$L_q^* = \frac{\lambda^2\sigma^2 + \rho^2}{2(1 - \rho)} \tag{23}$$

$$W_q^* = \frac{L_q}{\lambda} \tag{24}$$

4. Numerical Results

We present some numerical results (using SageMath, www.sagemath.org) for the modified (workload-dependent) M/G/1 queueing model from section 3 and compare them to results from a standard (constant service time) M/G/1 and a developed simulation (using Simul8, www.simul8.com).

Based on a slightly simplified version of Figure 1, we consider a server that is able to switch between two-speeds of service distributions. From our empirical study we calculate the associated parameters to be:

- Arrivals follow a Poisson process with mean inter-arrival time of 92 minutes and with an overall mean service time of 84 minutes, thus $\lambda = 1/92, \mu = 1/84$ and $\rho = 84/92 = 0.91$. Note that from the data we have estimated the demand per single member of staff. Typically there is a fixed number of staff on the daytime shift, so it is possible to derive demand rates per single server.
- We observe from Figure 1 a possible switching threshold of around 7 patients, so initially we take $A = B = 6$ i.e. when 6 or less patients are waiting we use service speed 1 (slow), and for 7 or more patients waiting we use service speed 2 (fast).
- Service speed 1 follows a lognormal distribution with mean of 86 minutes and variance of 77. Thus $\mu_1 = 1/86, \sigma_1^2 = 77, \rho_1 = 86/92 = 0.93$
- Service speed 2 follows a lognormal distribution with mean of 62 minutes and variance of 55. Thus $\mu_2 = 1/62, \sigma_2^2 = 55, \rho_2 = 62/92 = 0.64$
- For our simulation model, we use exactly the same distributions and parameter as those above.
- From our data, after adjusting for multiple staff on roster, we estimate that the observed mean number of patients waiting (for a single server) is 2.4 patients with a mean waiting time of 221 minutes.

Table 1 shows how our three models compare: *constant server*, *modified server*, and *simulation*, corresponding in turn to the standard M/G/1 queue, the modified workload-dependent M/G/1 queue, and the simulation model. Results for the constant server are found using equations (23) and (24), and for the modified server equations (21) and (22). The simulation model has been run for 1,000 repetitions, each one for 100 shifts (of 8 hours duration) and with a 100 shifts warm-up period. These parameters were found to be more than adequate to give sufficient precision in the model predictions.

Table 1: Numerical results for mean queue length and waiting time for service

Model	**Mean queue length, L_q**	**Mean waiting time (mins), W_q**
Constant server	3.57	334
Modified server	2.86	261
Simulation	2.91	270
Observed	2.43	221

From Table 1 we immediately note that the modified server, accounting for the adaptive service speed, provides similar results to both the simulation model and that observed in the empirical study for mean queue length, although both slightly overestimate the observed mean waiting time. The constant server overestimates system congestion, indicating that it fails to fully capture the subtleties of an adaptive service rate and a workforce that our empirical study has shown speeds up to respond to the workload.

We now examine different switching thresholds. Figures 3 and 4 show results for equations (21) and (22) respectively, for different levels of switching thresholds where for now we again assume $A = B$.

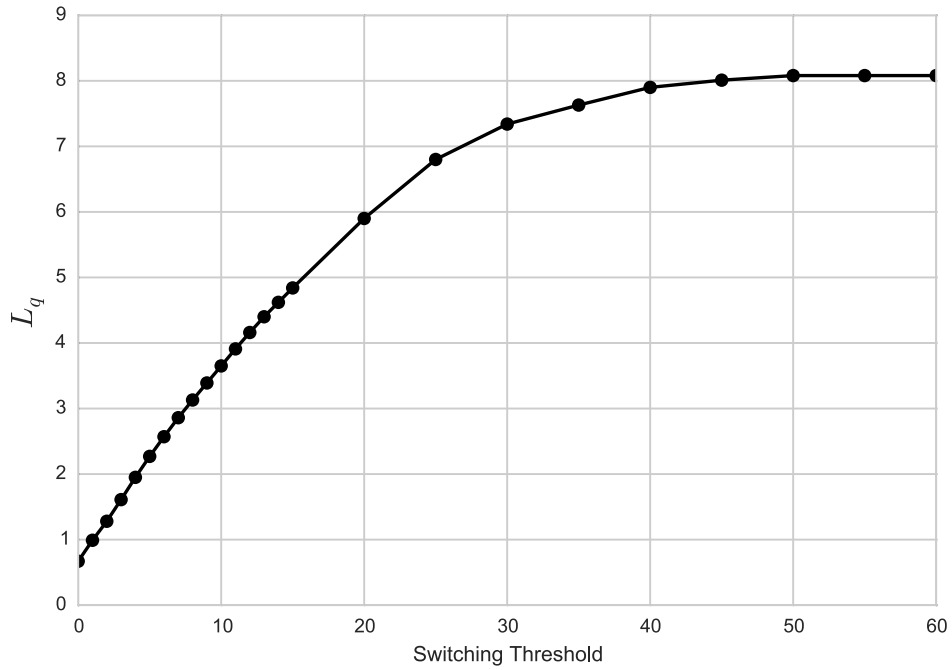

Figure 3: Plot of mean queue length L_q for different values of the switching threshold (where A and B are equal)

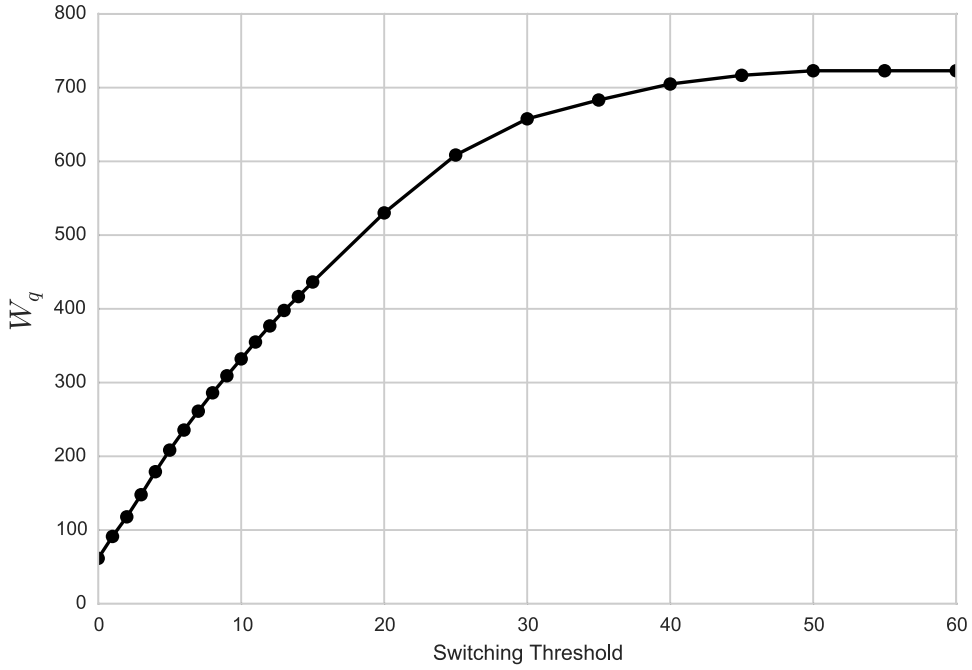

Figure 4: Plot of mean waiting time for service W_q for different values of the switching threshold (where A and B are equal)

Both Figures 3 and 4 illustrate how the adaptive service model could be useful to consider systems where staff behaviours and changing service rates impact on patient throughput and waiting times. The mean waiting time for service ranges from 62 minutes (where staff always work at the higher speed) to 723 minutes (where staff always work at the slower speed). These results are quite different than simply assuming the average service rate in a fixed speed server queueing model. Thus the immediate implication of this study is that the adaptive behaviour of the healthcare workforce will result in potentially very different levels of system performance than one might have otherwise predicted if assuming a fixed single service time distribution.

Finally, to illustrate the use and insights from the developed adaptive service model, we allow for different threshold values for A and B. Recall that in our model an adaptive server will change speed when the number of patients in the ED reaches B and will retain this speed until such a time when the number of patients waiting reduces below A ($0 \leq A \leq B$). This may be particularly useful in circumstances where staff continue to work faster until the backlog of patients is cleared and the number still waiting for service is below the original threshold. Figures 5 and 6 show how the mean number and waiting time in the queue and varies over a range of different values of A and B.

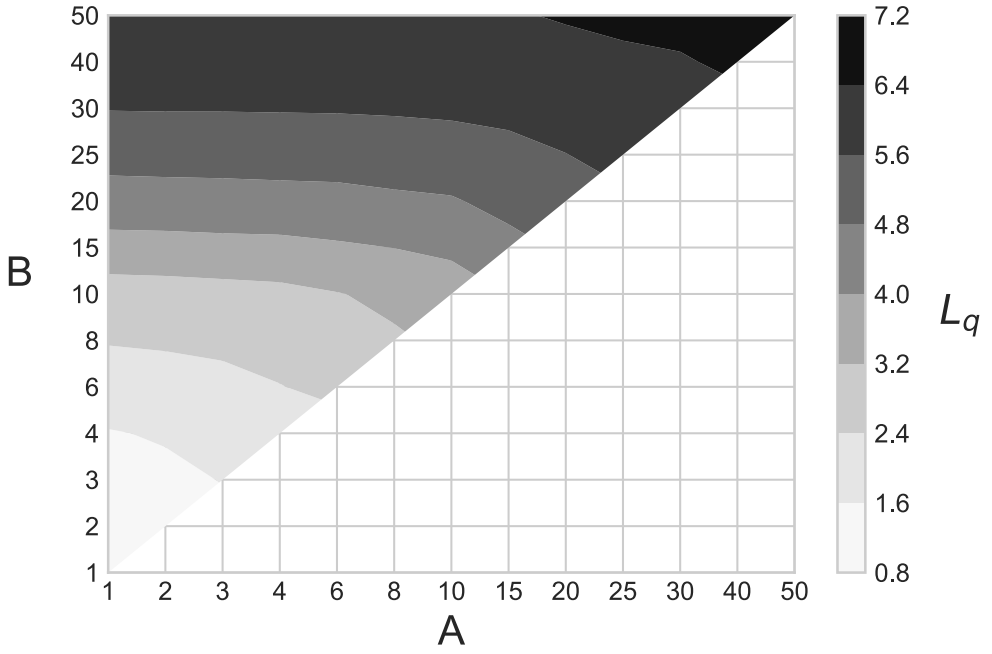

Figure 5: Contour plot of mean queue length L_q for different values of the switching threshold values A and B.

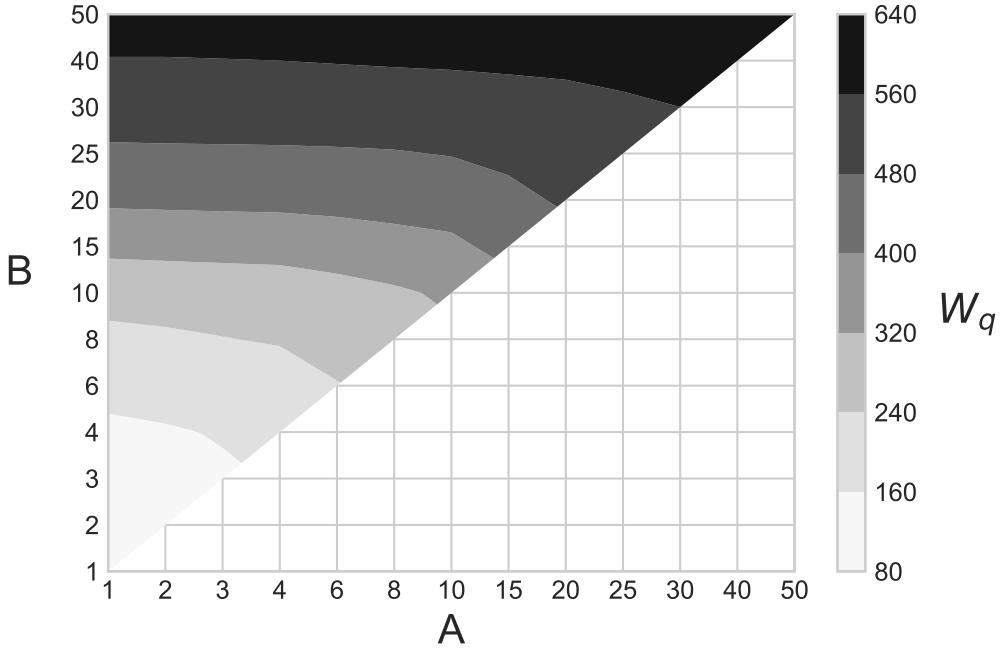

Figure 6: Contour plot of mean waiting time W_q for different values of the switching threshold values A and B.

In Table 1 we assumed that the switching thresholds A and B were fixed and equal ($A = B = 6$). The modified server and simulation however still slightly over-estimated the observed mean waiting time for service. On closer examination of the numerical results used to construct Figure 6, we might hypothesise that in fact in the ED under investigation staff move to the faster speed of service for $B = 6$ and continue to work at the faster speed until the number of patients waiting drops below 4 (i.e. $A = 4$). With these parameter values, we obtain a similar waiting time to that observed (221 minutes). From our empirical study it is impossible to know for sure the threshold values, but using the developed modified server model has the additional benefit of allowing us to estimate what these values might actually be in practice.

5. Workload Fatigue and Service Breakdown

Having explored staff behaviours in relation to service speed and workload, we now turn our attention to the related issue of workload fatigue and service breakdown. Again it is stressed that whilst the new model proposed here is still a simplified version of reality, it nevertheless captures new behavioural features not typically considered in the literature. Furthermore, this second model is presented to provide an additional and complementary motivating example of healthcare behavioural queueing theory and to motivate further work in this important and emerging field of research.

There is a sizeable medical literature on staff burnout caused by prolonged and sustained high levels of demand for care and resulting workload. Of concern is the reported detrimental impact not only on the healthcare workforce (such as a reduction in job satisfaction, increased sickness rates and staff turnover) but also on patient outcomes; see for example Gaba and Howard (2002), Coomber and Barriball (2007), You *et al* (2013), and Rubulotta *et al* (2016).

In this paper, we propose to capture this server behaviour through a queueing system model which incorporates service breakdown resulting in patients being released from that resource and having to be treated by other staff or transferred to a different healthcare setting. It could be viewed that staff burnout whilst on shift could be captured by server breakdown models such as service with vacation models (see for example Gray *et al*., 2000). Here we instead desire to add to the literature, and thus extend the choice of available models dependent on the particular situation at hand, by proposing a variant on the well-studied classical machine breakdown literature.

We also note that capturing such a system might equally lend itself to modelling of pre-emption by higher acuity patients (such as urgent and critical cases). That is, staff may become unavailable to treat more critical patients and the care for lower acuity cases could be impacted.

In considering a member of staff, such as a specialist consultant in a busy ED, we allow the possibility that given a prolonged period of high workload and faltering productivity (as evidenced in the case study in Section 2 and related literature),

the member of staff is taken off-shift for a much needed rest break before returning after a given time, or a replacement member of staff is sought to replace them. During this time, no patients are assigned to them and are instead moved to another member of staff (equivalently, in a pre-emptive situation, a member of staff may be required to provide care for critical patients and thus leave lower acuity patients). If this is a specialist care provider, where no other local workforce can provide similar care, patients would need to be moved to an alternative care setting until such a time that the staff member comes back on shift or an alternative member of staff to cover the breakdown is called in. In the later case, during that time no further patients are allowed to enter the system and are diverted to the alternative provider. We call this system *catastrophic* service breakdown, to distinguish it from a server with vacation. Such a model would of course also be applicable to modelling a wide range of catastrophic events, such as an ED shutdown to the general public in order to prioritise and cope with a major incident such as serious road traffic accident or terrorist attack.

Consider a single server queue where patients arrive according to a Poisson process with mean arrival rate λ and with service times following an exponential distribution with a mean $1/\mu$. Arrival and service times are independent of each other. Patients are served on a FCFS basis. The server is 'on' for a random time distributed exponentially with mean $1/\alpha$ after which a catastrophic event occurs (the member of staff leaves). The server stays away for a random time distributed exponentially with mean $1/\beta$. We model the system as a continuous time Markov chain (Figure 7) where $X(t) = i$ (for $i = 1, 2, 3, \ldots$) when there are i patients waiting for service in the system and the server (staff member) is on (available to serve or busy working) at time t. In addition, let $X(t) = D$ denote that the server is down (e.g. suffers workload burnout and is unavailable) at time t.

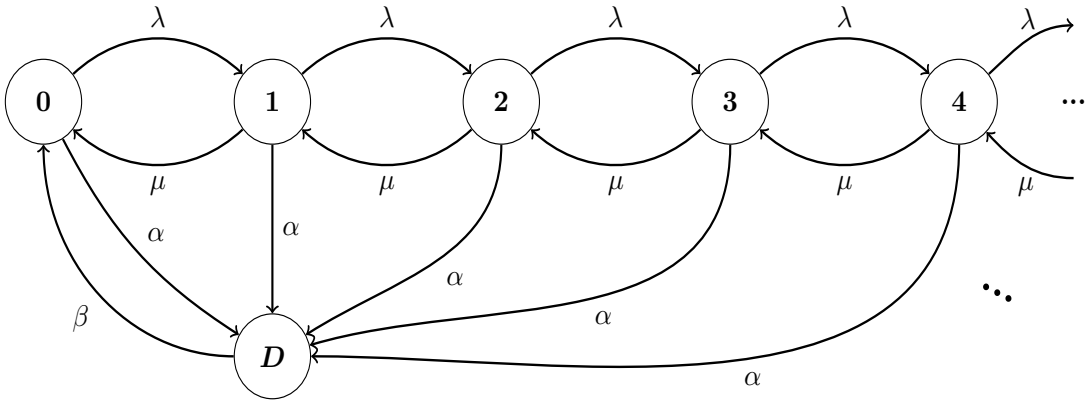

Figure 7: Markov chain for the catastrophic service breakdown model

For $j = D, 0, 1, 2, \ldots$ let $p_j = \lim_{t\to\infty} P\,[X(t) = j]$. To obtain steady-state probabilities p_j, consider the following balance equations:

$$\alpha(p_0 + p_1 + \cdots) = \beta p_D$$

$$\beta p_D + \mu p_1 = (\lambda + \alpha)p_0$$

$$\mu p_2 + \lambda p_0 = (\lambda + \alpha + \mu)p_1$$

$$\mu p_3 + \lambda p_1 = (\lambda + \alpha + \mu)p_2$$

$$\mu p_4 + \lambda p_2 = (\lambda + \alpha + \mu)p_3$$

$$\vdots$$

From the first equation we have $p_D = \alpha/(\alpha + \beta)$ since $p_0 + p_1 + \cdots = 1 - p_D$. Multiplying the second equation by 1, third by z, fourth by z^2 and so on, and summing up with obtain:

$$\beta p_D + \frac{\mu(\psi(z) - p_0)}{z} + \lambda z\psi(z) = (\lambda + \alpha + \mu)\psi(z) - \mu p_0$$

where $\psi(z) = p_0 + p_1 z + p_2 z^2 + p_3 z^3 + \cdots$ We note that unlike typical moment generating functions, here $\psi(1) = 1 - p_D$. Rearranging we obtain:

$$\psi(z) = \frac{\mu p_0 - z\beta p_D - p_0 \mu z}{\mu + \lambda z^2 - \lambda z - \alpha z - \mu z} \qquad (25)$$

The only unknown in (25) is p_0. However standard approaches such as $\psi(0) = p_0$ and $\psi(1) = \beta/(\alpha + \beta)$ do not yield a solution for p_0. We note that $\psi(z)$ is a continuous, differentiable, bounded, and increasing function over $z \in [0,1]$ and from (25) is of the form $\phi(z) = A(z)/B(z)$ where $A(z)$ and $B(z)$are polynomials corresponding to the numerator and denominator of the equation. If there exists $z^* \in [0,1]$ such that $B(z^*) = 0$, then $A(z^*) = 0$ otherwise it violates the condition that $\psi(z)$ is a bounded and increasing function over $z \in [0,1]$. By setting the denominator of $\psi(z)$ in (25) to zero, we obtain:

$$z^* = \frac{(\lambda + \mu + \alpha) - \sqrt{(\lambda + \mu + \alpha)^2 - 4\lambda\mu}}{2\lambda}$$

Setting the numerator of $\psi(z)$ in (25) to zero, we get:

$$p_0 = \frac{\alpha\beta z^*}{(\alpha + \beta)\mu(1 - z^*)}$$

By substituting for z^* we obtain p_0 as:

$$p_0 = \frac{\alpha\beta}{\mu(\alpha+\beta)}\left[\frac{\lambda+\mu+\alpha-\sqrt{(\lambda+\mu+\alpha)^2-4\lambda\mu}}{\lambda-\mu-\alpha+\sqrt{(\lambda+\mu+\alpha)^2-4\lambda\mu}}\right] \quad (26)$$

Furthermore, by rearranging terms in (25), we get the function $\psi(z)$ as:

$$\psi(z) = \frac{\alpha p_0(1-z) - z\alpha\beta/(\alpha+\beta)}{\lambda z^2 - (\lambda+\mu+\alpha)z + \mu} \quad (27)$$

We now derive some steady-state performance measures. Let P_l be the probability that a patient is lost (moved to another server or healthcare setting) and W be the average response (or sojourn) time for patients that are served. Let L be the time-averaged number of requests for service in the system in the long run (note that it includes the downtimes when there are no requests in the system). By definition:

$$L = 0p_D + 0p_0 + 1p_1 + 2p_2 + 3p_3 + \cdots$$

and that can be written as $L = \psi'(1)$. By taking the denominator of $\psi(z)$ in (27) and letting $z = 1$, we get the average number of patient requests in the system as:

$$L = \frac{1}{\alpha}\left[\frac{\lambda\beta - \mu\beta + p_0\mu(\alpha+\beta)}{\alpha+\beta}\right]$$

The number of requests that are dropped per unit time in steady state is $\alpha(1p_1 + 2p_2 + 3p_3 + \cdots) = \alpha L$. Hence the fraction of requests that entered the queue and were dropped when the server left is $\alpha L/(\lambda(1-p_D)$. The probability that an arriving patient will complete service, given it arrived when the sever was on, is given by (conditioning on the number of requests seen upon arrival):

$$\sum_{j=0}^{\infty}\left(\frac{p_j}{1-p_D}\right)\left(\frac{\mu}{\mu+\alpha}\right)^{j+1} = \frac{\mu}{\mu+\alpha}\frac{1}{1-p_D}\psi\left(\frac{\mu}{\mu+\alpha}\right) = \frac{\mu}{1-p_D}\frac{\beta - p_0(\alpha+\beta)}{\lambda(\alpha+\beta)}$$

Therefore, the rate at which patients exit the queue for that staff member is $\mu\beta/(\alpha+\beta) - \mu p_0$. Since the drop rate (derived earlier) is αL, we can write $\mu(\beta/(\alpha+\beta)) - \mu p_0 = \lambda(1-p_D) - \alpha L$. Thus the loss probability is $(\lambda p_D + \alpha L)/\lambda$ and by substituting for p_D we obtain P_l in terms of L as:

$$P_l = \frac{\alpha L(\alpha+\beta) + \lambda\alpha}{\lambda(\alpha+\beta)} \quad (28)$$

Finally, using Little's law we can derive W in the following manner. The expected number of patient requests in the system, when the server is on, is $L/(1-p_D)$. In steady state, of these requests a fraction $(\lambda(1-p_D) - \alpha L)/\lambda(1-p_D)$ will only receive service. Therefore the average response time at the server,

as experienced by patients who do receive service, is given by $L/\lambda(1-p_D)^2$, which yields:

$$W = \frac{L(\alpha+\beta)^2}{\lambda\beta^2} \tag{29}$$

To illustrate the model, we perform some numerical examples based on the ED under consideration. Arrivals follow a Poisson process with mean inter-arrival time of 92 minutes and with an overall mean service time of 84 minutes, so $\lambda = 1/92, \mu = 1/84$. From our empirical study (Figures 1 and 2) for the highest levels of workload (in excess of 11 patients waiting), there is an apparent increase again in service times, or conversely a drop-off in staff productivity. From our ED data, we compute that the proportion of time when the number waiting for service per server exceeds 11 patients is 1.8%. In turn this equates to, on average, a period of 940 minutes between successive occasions where 12 or more patients are waiting. Hence to illustrate the use of the catastrophic breakdown of service model, we take $\alpha = 1/940$ and with β taking a range of values between 100 and 400 minutes.

Figures 8 and 9 show the corresponding plots of L (average number of patients in the system) and W (average patient sojourn time) respectively. They help show that as the duration of time that the server is off increases (i.e. β increases), the number of patients able to be served by that member of staff decreases, and thus as the system becomes less congested sojourn times also decrease. However this is at the expense of increased numbers of patients lost to the system. The relationship β, L and W is however non-linear as clearly shown in both plots.

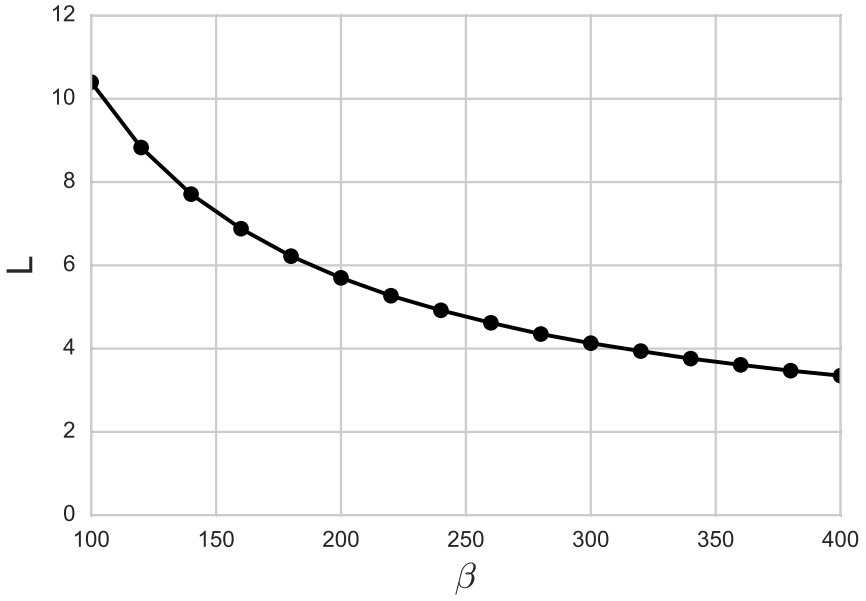

Figure 8: Plot of mean number of patients in the system, L, for different values of β (with $\lambda = 1/92, \mu = 1/84$ and $\alpha = 1/940$)

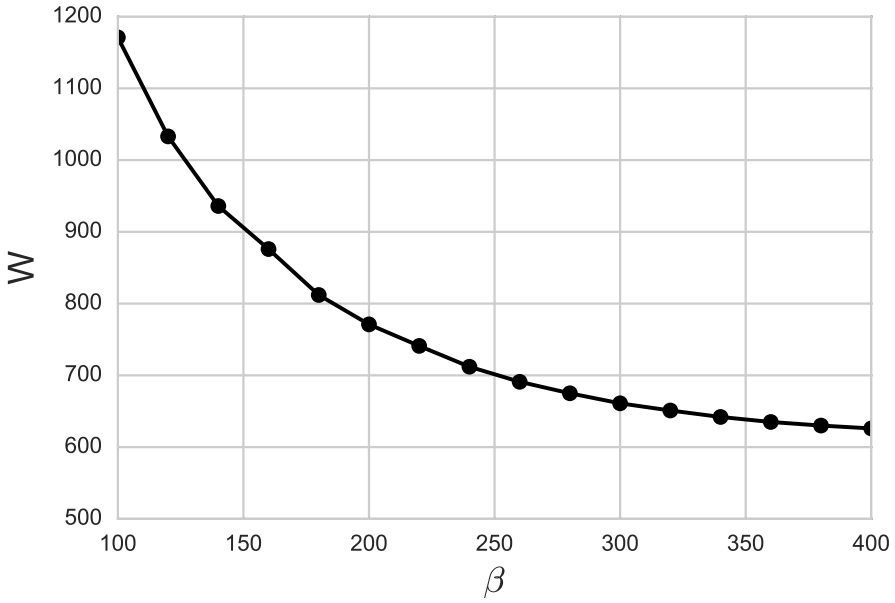

Figure 9: Plot of mean patient response (sojourn) time, W, for different values of β (with $\lambda = 1/92, \mu = 1/84$ and $\alpha = 1/940$)

6. Conclusions

Motivated by an empirical study of staff behaviours in an Emergency Department (ED), this paper builds on the literature to more formally consider queueing theory models for explicit consideration of situations when the time it takes a resource to serve for a patient depends on the current state of that queueing system, specifically the workload as measured by the current queue length for service. Our study indicates that service times decrease with an increase in workload. By staff working faster, the ED increases its throughput when it is busy. The implications are that the adaptive behaviour of the healthcare workforce increases the overall process flow of patients that one might have not have otherwise observed if assuming a fixed single service time distribution.

Motivated by our empirical study and similar previously published findings, an analytical queueing model, namely an M/G/1-type model has been considered, which is able to capture some of the observed server behaviours as opposed to classical queueing theory models with assumed constant service rates. The model allows for two-speeds of service, with thresholds (numbers of patients waiting for service) at which the service rate switches. From our numerical results, the immediate implication of this paper is that the adaptive behaviour of the healthcare workforce will result in potentially different levels of system performance than one might have otherwise predicted if assuming a fixed single service time distribution.

Having explored staff behaviours in relation to service speed and workload, we then considered the related issue of workload fatigue and service breakdown.

Indeed there is a sizeable medical literature on staff burnout caused by prolonged and sustained high levels of demand for care and resulting workload. In this paper, we capture this server behaviour through a queueing system model which incorporates service breakdown resulting in patients being released from that resource and having to be treated by other staff or transferred to a different healthcare setting. We call this system *catastrophic* service breakdown, to distinguish it from a server with vacation. We illustrate use of the catastrophic service breakdown model applied to the empirical study for the ED under investigation.

The research presented in this paper helps to demonstrate the importance of more accurately capturing server behaviours in workload-dependent environments, and the impact this has on the overall system performance. It is hoped that this paper might help spawn an emergence of behavioural queueing systems literature, retaining the role that queueing theory plays within our field but with enhanced consideration of behaviours when constructing such models. More immediate natural extensions to this work would be to consider the development of analytical models with multiple switching thresholds covering a range of underlying distributions and assumptions, such as prior related work by Zhernovyi (2012) and Baër *et* al. (2014), and to allow α in the catastrophic model to be a function the number of patients waiting for service i. Furthermore, it would be useful to examine the generalizability of our findings by studying other EDs, or indeed across different healthcare settings, to ascertain whether workforce adaptive behaviour is a common feature and to begin to understand factors that influence switching thresholds and productivity. To help progress this, future planned work will include academic colleagues with expertise in ethnography.

Acknowledgements

The author is grateful to the helpful suggestions made the anonymous reviewers, and to feedback received from colleagues at conference presentations of this research, including INFORMS Annual Meeting (Houston, 2017) and OR60 (Lancaster, 2018).

References

Aiekn LH, Clarke SP, Sloane DM, Sochalski J and Silber JH (2002), "Hospital nurse staff and patient mortality; nurse burnout and job dissatisfaction". *Journal of the American Medical Association*, 288: 1987-1993.

Anand KS, Paç MF and Veeraraghavan S (2011), Quality–Speed Conundrum: Trade-offs in Customer-Intensive Services. *Management Science* 57(1): 40-56

Baër N, Boucherie RJ and van Ommeren JCW (2014), The *PH*/*PH*/1 multi-threshold queue. In *Analytical and Stochastic Modeling Techniques and Applications, Proceedings of the 21st International Conference, ASMTA 2014*. Springer, 95-109.

Barnes S, Golden B and Price S (2013), "Applications of Agent-Based Modeling and Simulation to Healthcare Operations Management". In *Handbook of Healthcare Operations Management,* Volume 184 of the series International Series in Operations Research & Management Science, Springer, Chapter 3: 45-74.

Brailsford SC, Harper PR and Sykes J (2012), "Incorporating Human Behaviour in Simulation Models of Screening for Breast Cancer". *European Journal of Operational Research*, 219: 491-507.

Brailsford SC, Harper PR, Patel B, and Pitt M (2009) "An Analysis of the Academic Literature on Simulation and Modeling in Healthcare". *Journal of Simulation.* 3: 130-140.

Coomber B and Barriball KL (2007), "Impact of job satisfaction components on intent to leave and turnover for hospital-based nurses: A review of the research literature". *International Journal of Nursing Studies*, 44(2): 297–314.

Faddy M, Graves N and Pettitt A (2009). Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2): 309-14.

Fetta AG, Harper PR, Knight VA, Vieira IT and Williams JE (2012), "On the Peter Principle: An agent based investigation into the consequential effects of social networks and behavioural factors". Physica A, 391: 2898–2910.

Franco LA and Hämäläinen RP (2016), "Behavioural operational research: Returning to the roots of the OR profession". *European Journal of Operational Research*, 249 (3): 791-795.

Gaba DM and Howard SK (2002), "Fatigue Among Clinicians and The Safety of Patients". *New England Journal of Medicine*, 347(16): 1249-1255.

Garg RL and Singh P (1993), "Queue-Dependent servers Queueing System". *Microelectronics Reliability*, 33(15): 2289- 2295.

Gebhard RF (1967), "A Queueing Process with Bilevel Hysteretic Service-Rate Control". *Naval Research Logistics Quarterly*, 14(1): 55-67.

Gray WJ and Wang P (1992), "An M/G/1-type queuing model with service times depending on queue length". *Applied Mathematical Modelling*, 16: 652-658.

Gray WJ, Wang PP and Scott M (2000), "A vacation queueing model with service breakdowns". *Applied Mathematical Modelling*, 24: 391–400.

Hämäläinena, RP Luomab J and Saarinen E (2013), "On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems". *European Journal of Operational Research*, 228 (3): 623–634.

Hulshof PJH, Kortbeek N, Boucherie RJ, Hans EW and Bakker PJM (2012), "Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS". *Health Systems*, 1(2): 129-175

Jaeker J and Tucker A (2016), "Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* 63(4): 1042-1062.

Kc D and Terwiesch C (2009), "Impact of workload on service time and patient safety: An econometric analysis of hospital operations". *Management Science* 55(9): 1486–1498.

Kc D and Terwiesch C (2013), "An Econometric Analysis of Patient Flows in the Cardiac Intensive Care Unit". *Manufacturing & Service Operations Management,* 14(1): 50-65.

Knight VA and Harper PR (2013), "Selfish routing in public services". *European Journal of Operational Research*. 230 (1): 122–132.

Lin CH and Ke JC (2011), "Optimization Analysis for an Infinite Capacity Queueing System with Multiple Queue- Dependent Servers: Genetic Algorithm". *International Journal of Computer Mathematics*, 88(7):1430-1442.

Rubulotta F, Scales DC and Halpern SD (2016), "Night shifts, human factors, and errors in the ICU: a causal pathway?" *Intensive Care Medicine* 42: 456-457.

Saaty TL (1961), *Elementary of Queueing Theory with Applications*. McGraw-Hill, New York.

Tan TF (2014), "When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity". *Management Science*, 60(6): 1574–1593.

Tirdad A, Grassmann WK and Tavakolia J (2016), "Optimal policies of M(t)/M/c/c queues with two different levels of servers". *European Journal of Operational Research*, 249(3): 1124-1130.

Wang KH and Tai KY (2000), "A Queueing System with Queue-Dependent Servers and Finite Capacity". *Applied Mathematical Modelling*, 24(11): 807-814.

You L, Aiken LH, Sloane DM, Liu K, He G, Hu Y, Jiang X, Li X, Li X, Liu H, Shang S, Kutney-Lee A, Sermeus W (2013), "Hospital nursing, care quality, and patient satisfaction: Cross-sectional surveys of nurses and patients in hospitals in China and Europe". *International Journal of Nursing*, 50(2): 154–161.

Zhernovyi YV (2012), "Stationary Characteristics of $MX/M/1$ Systems with Two-Speed Service". *Journal of Communications Technology and Electronics*, 57(8): 920-931.