The role of queuing models in economic evaluation: lessons from an illustrative case study of emergency care

Eduardo Gonzalo-Almorox
Alexandru Lucian Gaianu
Luke Vale

October 2015

Working paper reference number
HEG2015_03

Eduardo Gonzalo Almorox,
Health Economics Group
Institute of Health & Society,
Newcastle University
Baddiley-Clark Building,
Richardson Road,
Newcastle Upon Tyne, NE2 4AX.
Email: e.gonzalo-almorox@newcastle.ac.uk

Disclaimer:
The Health Economics Group discussion paper series promotes the dissemination of recent work and work in progress to a wider audience. The work and ideas reported here do not represent the final position and need to be treated as work in progress. The material and views expressed in the series are solely those of the authors and should not be quoted without their permission. Comments are welcome, and should be sent to the corresponding author
The role of queuing models in economic evaluation: lessons from an illustrative case study of emergency care

Eduardo Gonzalo-Almorox¹
Alexandru Lucian Gaianu²
Luke Vale³

1. NIHR Research Methods Fellow, Health Economics Group, Institute of Health & Society, Newcastle University.
2. NIHR Research Methods Intern, Health Economics Group, Institute of Health & Society, Newcastle University
3. Health Foundation Chair in Health Economics, Health Economics Group, Institute of Health & Society, Newcastle University.

Address for correspondence:

Eduardo Gonzalo Almorox,
Health Economics Group
Institute of Health & Society,
Newcastle University
Baddiley-Clark Building,
Richardson Road,
Newcastle Upon Tyne, NE2 4AX.

Email: e.gonzalo-almorox@newcastle.ac.uk

October 2015
The role of queuing models in economic evaluation: lessons from an illustrative case study of emergency care

Abstract

This paper discusses the suitability of queuing theory as a component in economic evaluations. Receiving timely acute care is of major concern to patients, health care providers, and policy-makers. Underlying capacity constraints associated with the provision of care might be neglected in economic evaluations. When they are considered, the type of models usually employed are both data and computationally demanding. We argue that queuing modelling may be an appropriate framework to use within economic evaluation when capacity constraints play an important role. We illustrate the role of queuing theory using a dataset with 33,926 records, corresponding to one year activity of the ambulance service from the North East of England. We use these data to explore alternative scenarios about the ways in which services can be centralised. The results of the model are expressed in terms of waiting times and allow for the identification of possible inefficiencies within a system. Results may also inform the design of potential policies for the reorganisation of emergency care which take into account the resource constraints analysed.

Keywords: Economic evaluation; queuing theory; organisation of emergency care; centralisation of health services.

1 The authors want to acknowledge the valuable comments from Katherine Payne, Alec Norton and Peter McMeekin on early versions of this manuscript as well as all the participants in the I EuHEA Conference for Early Careers and PhD students and seminars at the Newcastle Health Economics Group. All errors are authors’ responsibility. This research has been funded by a NIHR Research Methods Fellowship programme.
1. Introduction

Emergency care in the UK includes paramedics who in most circumstances operate outside of hospitals, Accident and Emergency (A&E) units and emergency wards of hospitals designed to accept emergency admissions (Boyle, 2011). The organisation of emergency care relies on the timely coordination of events along the pathway. A successful and timely management of the service is determined by the quantity of resources available as well as when and where these resources are available. Resource availability, which enables shorter time to treatment, determines the feasibility of certain therapies and hence their effectiveness and efficiency (De Luca et al., 2004; Gumbinger et al., 2014; Penaloza-Ramos et al., 2014). Likewise, delays may affect the subsequent use of health care resources. Huang et al. (2010), showed that patients who experienced delays to treatment in an acute ward were associated with an 11% higher total costs and a 12% longer inpatient length of stay compared to those who received treatment in a timely manner. A successful and prompt provision of emergency care is therefore determined by good coordination between the ambulance service and A&E units to ensure that the patient arrives at the right time to the appropriate centre that provides the care s/he requires.

The decision to refer patients to a particular medical centre is challenging. Ambulance staff need to balance the pay-offs associated with the decision of taking patients to the nearest health centre which also minimises the time to treatment as well as the need to get back to service as soon as possible, against transferring the patient to a further but better equipped centre. Although a centre may be closer to the event, it may not have the most appropriate level of expertise and facilities. Going to a better equipped and more experienced centre may require the ambulance to be dedicated (because of the extra time to travel to and from the patients location) for longer. Considering the management of urgent episodes, there is equivocal evidence regarding the most cost-effective care strategy. Pons et al. (2005) reported a positive association between the reduction in response times (as an indicator of the ambulance service performance) and the achievement of better survival for cardiac conditions. Wilde (2013) extended this analysis showing that response times of an ambulance service also predicted the incidence of additional adverse effects not necessarily linked to the original health condition. Other researchers have explored the consequences of moving patients to more specialised centres and hence incurring in greater travel times. Hunter et al. (2013) compared the effects of a strategy based on centralising acute stroke care services. At a 90 day time horizon, their model predicted better survival rates and the cost savings after incorporating the strategy in London. On the same basis, Morris et al. (2014) analysed alternative health outcomes such as
the risk adjusted mortality and the length of hospital for similar settings showing improved outcomes associated with a reconfigurations of services. Similarly, McMeekin et al. (2013) examined the implications associated with the relocation of patients from local to regional services in terms of variations in their resource use. In their study they include variations of the resource use during the pre-hospital care (e.g. extra ambulance journey), and also at the hospital (e.g. increase in the capacity and the costs of management of bigger centres) and subsequent phases after discharge (e.g. social care required). Albers et al. (2012) also reviewed the economic effects related to the centralisation of cancer services and found inconclusive evidence about the cost-effectiveness of centralising cancer services. Thus, although delaying a treatment may reduce outcomes, a delayed but more appropriate treatment may improve them.

In the UK the 999 service manages the dispatch of ambulances. It has experienced a dramatic increase in the number of calls received from 4.7 million in 2001/02 to about 9 million in 2012 (Fernandes, 2011; Keogh, 2013). Similarly, admissions to A&E units have risen by 50% since 2007-2008. These increases in demand, in addition to turnovers in staff and financial constraints, make the National Health Service (NHS) particularly concerned with the organization the early phases of care for acute events (Blunt et al., 2015). In 2013, Keogh (2013) proposed a centralisation of the current emergency services structure. The existing system consists of different types of facilities under a common umbrella defined as A&E. These may not cover the same events nor fulfil the same needs, though. The core rationale of the 2013 proposal implies grouping A&E centres into two broad groups. The first category, defined as emergency centres, will cover all type of patients by assessing and initiating their treatment. They would be placed on a local basis. Additionally, the second main category, labelled as major emergency centres, would centralise in fewer centres placed strategically those activities that required more specialized care and will receive a greater amount of resources.

It is argued that a concentration of health services leads to efficiency gains derived from the economies of scale and scope linked to the concentration of services (Goddard and Ferguson, 1997). Such a strategy therefore entails a variation of the existing capacity constraints. Yet, although these boundaries constitute a key element for the organisation and management of health care, many models still ignore them either implicitly or explicitly as well as their implications for the economic evaluations. We propose a queuing model as an approach for
use within economic evaluations which overcomes this limitation. Our rationale is that by using
a queuing model, we may improve the design of an economic evaluation by reflecting more
realistically and flexibly different care strategies than is readily possible in transition models
but in a way that is less complex to implement than discrete event simulations (DES). The latter
are appealing as they can represent the implications of resource limitations (Karnon and Afzali,
2014). Yet, DES are data demanding and normally require highly specialised expertise to
populate, validate and implement the models in use (Brennan et al., 2006).

A queuing model describes a system with a flow of entities that arrive to one or more servers
in order to receive a certain service. Servers can be either busy executing a task, or conversely
be free and ready to offer a service. Various economic questions have been addressed using
queuing models. These include the valuation of systems considering the cost of delay and the
consequent design of pricing strategies for customers (Mendelson, 1985; Naor, 1969), the
effect of time delays on the competitive positioning of different service providers and the effect
of time delays on product quality within market structures where companies compete in terms
of time (De Vany and Saving, 1983; Levhari and Luski, 1978). Health care frameworks have
benefited from queuing methods for estimating waiting times, analysing the utilization of
health resources and designing appointment systems (Fomundam and Herrmann, 2007).
Hitherto, the potential to apply this approach within economic evaluation has not received
much attention.

Given the implications of the resources employed for delivering emergency care, this paper
shows how queuing theory could be used for planning and organising emergency care services.
We compare and assess the possible underlying time costs spent delivering care under different
scenarios. Each of these scenarios entails a different set of resources (e.g. workforce and/or
facilities for providing a particular care), location of the centres and composition of the demand
that should receive a specific care. All these issues require a careful analysis for the design of
a subsequent economic evaluation. By studying the flows of patients within the pre-
hospitalisation pathway, this paper sheds light on how potential resource constraints may affect
the performance of the system. We do so by measuring the time lost and gained under different
hypothetical care structures (i.e. local versus centralised care). We contribute to the current
literature on economic evaluation by addressing two goals. First we show how a queuing model
can be used analysing the performance of an emergency ambulance unit. This performance is
assessed on the basis of average time taken to receive the service. Secondly, we draw out the
methodological implications for conducting an economic evaluation when applying this modelling approach.

2. Methods

2.1 The ambulance emergency service

Emergency and urgent care services are typically characterised as a two-step process that involves ambulances and A&E units normally. The correct management of those episodes covered by the ambulance services and A&E units is critical given the potential spill-overs to other areas of care such as admissions to general hospitals\(^3\) (Hurst and Williams, 2012) or bed occupancy rates (National Audit Office, 2013). In the first step the ambulance service transports a patient from the event to the health centre. The second step is the provision of care in the health centre. We investigate solely the performance of the ambulance service in terms of its ability to appropriately transport patients to hospital. Figure 1 represents the main time components involved in a system of an ambulance service. The system begins with the reception of the call by the 999 service operator which evaluates the problem and, if the event requires a service, communicates this to the ambulance, which is subsequently dispatched to the incident. Capacity constraints, in the 999 operator service\(^4\) and in the availability of ambulances, may cause delays between the point when the patient attempts to contact the service and when the ambulance is dispatched. The following stage consists of the assessment and treatment of the patient at the place of the incident by the ambulance crew and/or other attending health care professional(s). Depending on the characteristics of each event, the time taken for this operation differs. Likewise, cases where the event does not require additional care, result in the return of the ambulance to the base (represented by dotted lines in Figure 1). The last phase of the care process concerns the transport of the patient to the medical centre. At this point there may be two possible alternative pathways to follow. Incidents categorised as mild severity cases may be referred to a local centre and receive initial care. Conversely, more severe cases, which might require specialised care, may be sent to a centre not necessarily located nearby. Finally, once the ambulance arrives to the medical centre and the patient is handed over to medical centre staff, the ambulance returns to its base site, unless it is called to attend a subsequent incident.

---

\(^3\) Admissions in general hospitals from A&E units represented a 67% of their total bed occupancy and implied an approximate cost £12.5 billion for the NHS during 2012-13.

\(^4\) In our illustrative analysis we do not include the possible constraints presented by the operators.
Other constraints may appear in the system. Particularly, some centres within the system may have limited available resources to tackle the complex cases. This may result in further delays due to extra time that has to be taken to transport the patient to a centre with appropriate treatment. Considering the context of the paper the latter may imply additional queue awaiting to use the ambulance service. Likewise, queues may be formed also in the second part of the pathway due to competing demands from other service users at receiving health centres. Yet, the focus of this paper is to illustrate the value of the queue approach for economic evaluation by focusing on the queue for the use of an ambulance service.

![Sequence of events in the ambulance service](image)

**Figure 1: Sequence of events in the ambulance service**

2.2 *Theoretical framework: queuing theory*

A queue model differentiates between entities which receive the service and servers that provide the service. In our model patients are assumed to be entities and the fleet ambulances servers that perform various activities\(^5\). Since we consider only one node, this model is a single queue.

---

\(^5\) The services may involve receiving the call, assessing the case, contact an ambulance and dispatch the ambulance.
From the perspective of a provider of a health service it may be interesting to know the number of patients that arrive to a health service at random points of time and the number of patients that the service is able to attend. Considering this, two parameters of interest are used for modelling the system as queue. The inter-arrival rate, or rate of arrival $\lambda$, is the expected number of arrivals to the system per unit of time. Analogously, the service rate, $\mu$, is the expected number of cases served per unit of time.

For the case of an ambulance service, the value of these parameters may vary temporally over the day (for instance day shifts versus night shifts) and between different months of the year (winter versus summer). There may be other issues which lead to constraints in the system. Within an emergency care context, an example of these may be expressed by the maximum number of patients who could wait prior to receive the service during the triage phase. The boundaries would be controlled by the urgency of cases and would imply variations in health outcomes whose impact would be reflected on QALY estimates. Capacity constraints may reflect the arrival intensity of the system (i.e. the number of arrivals) and the distributions used to characterise the flow of these arrivals. Likewise, the association between capacity constraints, which inform the length of a queue, and the arrival rate, distinguishes finite source models and infinite source (Boxma, 1986).

We estimate both $\lambda$ and $\mu$ from our available dataset and express them in minutes. $\lambda$ is represented by the number of calls received by the 999 service that actually require the dispatch of an ambulance. $\mu$ is given by the average number of emergencies solved per unit of time by the ambulance service (from the dispatch to the referral in an A&E unit). The model assumes that the arrival process is driven by a homogeneous Poisson process with constant rate $\lambda$. Hence, the time between successive emergency occurrences is exponentially distributed so that there is an underlying memoryless in the arrival process which implies that both the duration until a new entity enters in the system and the duration of the service provided to an entity already in the system, are independent of the time that has already passed (Ross, 1983). Thereby, the probability to call the 999 service is independent of the time taken by the ambulance to complete the service.

---

6 The most simplistic definitions of these models assume that clients could wait infinitely until the service is delivered without incurring any penalty. As we will see, the structure of the model does not include the implications of blocking or congestion in the system since clients are assumed to be served eventually.

7 See Data section for further details.

8 There are calls that are assessed by phone and do not require an ambulance service.

9 In case of non-homogeneous Poisson processes, $\lambda$ is a function of time.
We characterise the flow of entities throughout a First In First Out (FIFO) rule. Therefore, entities that arrive earlier to the server receive the service first irrespective to their condition. Other rules which might be within this framework are the processor sharing (PS)\(^\text{10}\) or the Last Come First Serve (LCFS) rule (Zonderland and Boucherie, 2012). The latter might be associated with the idea of some kind of prioritisation where some entities are assigned to the service ahead of others regardless of their arrival time; which is consistent with the triage principles used but is not further considered in this paper. Prioritisation also allow differentiation between pre-emptive and non-pre-emptive queues. In pre-emptive queues, the patient with the highest priority receives the service immediately regardless of the fact that other patients may be in the service already. Non-pre-emptive queues also discriminate for high priority customers letting them to jump to the beginning of the queue. However, unlike pre-emptive queues, they have to wait until the service is completed by the preceding customer before they can receive the service.

These models are expressed compactly as A/B/s/K/n\(^\text{11}\) (Kendall, 1953). Hence, \(A\) and \(B\) inform the distribution of the arrival times and the service times respectively. \(s\) indicates the number of servers, \(K\) is the capacity of the system so that includes the maximum number of entities in the system that are served and, \(n\) is the population size or number of patients in the system.

2.3 The model

As stated above, the model comprises a single queue to a unique node that represents the whole ambulance service. Patients enter into the system exogenously at a rate \(\lambda\) and are not subject to constraints in terms of waiting capacity so that there is no limit with respect to the number of patients who can call and wait prior to being assessed. This assumption neglects the unlikely circumstance of a significant constraint such as a limit in the number of calls handled by the operator and the subsequent withdrawal of patients from the system after some time waiting (by self-presenting to another provider or deciding no treatment is needed or the condition progressing to the point where treatment is no longer possible).

Various system performance variables which indicate how the resources are utilised and the number of patients who are receiving the service show the results of system. We calculate the service use, \(\rho\), also defined traffic intensity (Gross et al., 2013), by combining the inter-arrival

---

\(^\text{10}\) The jobs are partitioned in a manner that customers get simultaneously an even proportion of the server’s capacity. This rule in the case of an ambulance service would entail a shared attention by the ambulance crew.

\(^\text{11}\) This is known as the Kendall’s notation.
rate and the service rate. Hence, denoting the mean service time as $E(P)$ and taking into account the exponential distribution of the service, then we can establish the following:

$$\frac{\lambda}{s} E(P) = \rho \rightarrow \frac{\lambda}{(\mu s)} = \rho \quad (1)$$

Where (1) indicates the relative utilization of a given server (e.g. the ambulance service) at a given point in time and also whether the system is stable $\frac{\lambda}{s} E(P) < 1$ or not. Thus, the system is stable if the inter-arrival rate $\lambda$ is less than the service rate of all servers ($s$) combined. $\rho$ may also provide insights about the underlying economies of scale and scope associated with the ambulance service and inform the design of an economic evaluation. The perspective of the analysis and the nature of the problem under consideration determines the choice of the performance indicator. Thereby, given the relevance of time for emergency care we examine the overall time taken by ambulance service to respond and convey the patient. This approach is consistent with current practices where ambulance quality indicators are used to assess the quality of the service (see National Health Outcomes, Department of Health England (2012)).

The North East Ambulance Service (NEAS) defines its quality indicators on the basis of achieving specific time targets$^{12}$. The mean throughput time ($W$) is our performance variable of the system. Likewise, the sojourn time which represents the total time spent in the process of care ($T$) includes the time receiving the service ($S$) and the waiting time in the queue prior to receiving the service ($W_q$). In our analysis we examine different scenarios to explore how these throughput times vary.

$$T = S + W_q \rightarrow E[T] = W \quad (2)$$

Little’s formula (Little, 1961) associates the average length of the system ($L$), given by the average number of patients in the system, the sojourn time of the system and the arrival rate.

$$E[L] = \lambda W \quad (3)$$

(3) shows the equilibrium of the system where the average number of patients entering in the system equals to the number of patients leaving the system. This expression allows the definition of the PASTA$^{13}$ property which states that the mean number of jobs that arrive to the system at any instant of time is equal to the average length of the system so that the probability

---


$^{13}$ Poisson Arrivals See Times Averages.
of seeing the system in a particular state is the same as a stationary probability. Assuming a residual production time of $1/\mu$ and combining both Little’s formula and the PASTA property it is possible to estimate the mean throughput time in the system from the inter-arrival and service rates ($\lambda$ and $\mu$ respectively).

\[
E[L] = \lambda W \quad ; \quad W = E[L] \frac{1}{\mu} + \frac{1}{\mu}
\]

\[
W = \lambda W \frac{1}{\mu} + \frac{1}{\mu}
\]

\[
W = W \rho + \frac{1}{\mu}
\]

\[
W = \frac{1}{\mu(1 - \rho)} \quad (4)
\]

$W$ is a positive and differentiable function of $\lambda$ so that as service use $\rho$ tends to unity the average waiting time and thus the number of patients within the system tend to infinity.

On the basis of Keogh (2013) proposal, we assess the system’s performance according to two hypothetical scenarios with policy relevance in terms of dispatching strategies. The system consists of a $M/M/s$ queuing structure where $M$ indicates a queue with an arrival process characterised by a homogenous Poisson distribution and the service time is exponential. Nonetheless, in case the arrival and/or service process were governed by a general distribution (G) instead, M should be replaced by a G when characterising the system.

We incorporate two underlying types of resource constraints given by the number of ambulances available for the completion of the service and the number of health centres where patients can be dispatched. Under this configuration we consider $s$ servers able to provide the service. Scenario 1 reflects the current organisation of pre-hospital care. Scenario 2 is a hypothetical dispatching policy based on the centralisation of acute care services, which might be thought of as representing the situation where the services provided in a set of local hospitals are replaced with a smaller number of specialized units.

2.3.1 Scenario 1
Taking into account the three stages experienced by an ambulance service and illustrated in Figure 1, Scenario 1 is regarded as the baseline scenario and considers three possible destinations, Centre A, Centre B and Centre C, which provide local coverage but may not have specialised care. Under this Scenario 1, patients are transported to the nearest centre from the event. We estimate the arrival rate ($\lambda$) as the average time between arrivals for all the patients irrespective to their final destination. We follow a similar strategy for calculating the service rates ($\mu$). On the basis of Singer and Donoso (2008) we calculate the service rate of the ambulance service as the average of the service rate associated with each stage of the pathway. The service rate of the system, $\mu$, is calculated as the inverse of the sum of the particular average times associated with the stages involved in process: outbound ($\tilde{O}$), assessment ($\tilde{A}$) and inbound ($\tilde{I}$), referred to each server ($i$):

$$\tilde{\mu} = \frac{1}{\sum_{i=1}^{n} (O_i + A_i + I_i)} \quad (5)$$

2.3.2 Scenario 2

This scenario describes a new service organisation. Instead of three possible destinations, this model captures the effects of centralising activity into two health centres. This means that some events may be associated with greater travel distances to the health centre and therefore the time taken therefore treatment is received. By altering the time taken to reach the health centre then as we see in expression 5 altering the inbound time will affect the service rate and so the performance of the model.

Analogously to Scenario 1, we also assume that patients are dispatched to the closest centre in this new hypothetical configuration. The adjustment of the inbound time consists of varying the service rate corresponding to each server $i$ ($\mu_i$) associated with the inbound time by assuming that emergencies corresponding to the centre that is removed are re-directed to the nearest remaining centre (in terms of travel time). We compute the driving time from the particular location of the emergencies to the nearest centre and then re-calculate the new service rate parameter ($\mu_{i'}$). Likewise, the arrival rate ($\lambda$) remains constant.

3. Data and variables

We use a dataset composed of 33,926 episodes collected from the North East Ambulance Service (NEAS) National Health Service (NHS) Foundation Trust for a year (1st October 2009 to 30th September 2010). Our analysis includes admissions to an A&E unit in one of three
centres: A (2,102), B (10,991), C (10,531). We exclude 507, 2118 and 1707 observations corresponding to A, B and C respectively since they do not provide the necessary information for estimating the governing parameters (λ and μ). In this illustrative example we consider only a single ambulance station and a hypothetical available fleet of 10 ambulances at any point in time which is considered as the reference capacity.

Records report information on the time when the ambulance was contacted to approach the incident (Switch). We use these observations to estimate the inter-arrival rate (λ). In particular, λ is the average of the differences between subsequent time observations. The dataset also includes information regarding the time when the ambulance arrives at the incident (Arrive), leaves the incident (Leave) and the time of arrival to the medical centre (Hospital). Records are also referred to the location of the event – given by the postcode, the date of the event, a binary variable that indicates whether the referral of the patient to the health centre occurred during the day or overnight and a variable that categorises emergencies according to their severity. Hence, level A if for more severe cases whereas B corresponds to less severe cases.

For the purposes of our model we construct three additional variables to capture the time durations associated with different phases of the process represented in Figure 1. Our first variable, outbound journey reflects the time taken by an ambulance to arrive at the event once it is contacted by the 999 call centre. It is calculated as the difference between the time of arrival to the event and time when the ambulance is contacted. The second variable, Assessment, is the time taken by the ambulance crew to evaluate the incident. It is calculated as the difference between the time when the ambulances leaves the incident and the time when it arrives to the event. Finally, the third variable created is Inbound and represents the time taken by the ambulance to arrive at the hospital after the assessment of the patient is completed. It is the difference between the time recorded for a patient in the health centre and the time the ambulance leaves the event. These variables, included in expression (5), determine the service rate for each stage defined in the care pathway (e.g. arrival to the event, the assessment and the referral to the hospital). They also allow the estimation of the total time taken for each service.

3.1 Descriptive statistics

Figure 2 provides information about the flow of calls and the ambulance service. The top of the Figure 2 shows how activity referred to 999 call centre is distributed throughout the hours of the day and the caseload severity. The majority of cases that arrive at A&E centres are not severe. Nonetheless, severe cases are more likely to occur during the time range from 6:00 to
13:00 registering peak times at 8:00, 9:00 and 10:00. Likewise, the period from 8:00 to 13:00 seems to be the busiest in terms of total activity.

To test and verify that the arrival of patients into the system is effectively based on a non-stationary exponential process we divide our sample into three shifts, *Night* (from 0h00 to 7h00), *Morning* (from 8h00 to 15h00) and *Evening* (from 16h00 to 23h00), of 8 hours each. Then, we subsample these shifts into time intervals of 5 minutes and run a $\chi^2$ test in order to compare whether the records are distributed according to a Poisson distribution. All intervals are statistically significant at a 1% level of significance so that we can accept that the flow of calls is non-stationary distributed and effectively follows a Poisson process.

![Graph of Frequency of calls 999 service](image1)
![Graph of Journey duration](image2)

*Figure 2: Activity of the 999 call centre and the ambulance service*

The bottom part of Figure 2 shows the journey duration, both overall and split by the priority of the call. Journeys are most frequently between 30 and 40 minutes long (28%), with similar proportions of journeys taking either 20 to 30 minutes (20%) or 40 to 50 minutes (20%). The proportion of mild cases usually is greater than the proportion of severe.

Centres B and C receive similar proportions of the total demand (about 45% and 47% respectively) and A receives the remaining 9% of our effective sample. Table 1 shows the current activity with regards to each centre. This activity captured by showing the proportion of demand at each shift within for each centre and the time devoted for completing each stage.
Table 1: Descriptive statistics of each centre

<table>
<thead>
<tr>
<th>Health Centres</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Night</td>
<td>Morning</td>
<td>Evening</td>
</tr>
<tr>
<td>Proportion of</td>
<td>0.201</td>
<td>0.214</td>
<td>0.206</td>
</tr>
<tr>
<td>demand (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (minutes)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deviation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coeff.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variation</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Stage of the</td>
<td>Assessment</td>
<td>Assessment</td>
<td>Assessment</td>
</tr>
<tr>
<td>process</td>
<td>19.6</td>
<td>20</td>
<td>17.8</td>
</tr>
<tr>
<td>Mean (minutes)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>10.5</td>
<td>11.2</td>
<td>9.8</td>
</tr>
<tr>
<td>deviation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coeff.</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>Variation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (minutes)</td>
<td>16.9</td>
<td>18.5</td>
<td>9.9</td>
</tr>
<tr>
<td>Standard</td>
<td>11.3</td>
<td>14.9</td>
<td>5.2</td>
</tr>
<tr>
<td>deviation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coeff.</td>
<td>0.7</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Variation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In terms of the relative weight of activity given the three 8-hour shifts defined above, we see that the three selected A&E centres have similar levels of demand regarding each shift. Considering the centres uniquely, we can see that the A&E in A has a higher relative demand during the morning shift (45%) compared to C and B (both presenting a 41% of their respective activity). In B however, activity is similar across centres (for A 20% of total demand and 21% each for B and C).

Similarly, Table 1 provides information about the time duration of each stage in the process. Regardless of the centre of arrival, the assessment stage is the most time consuming part of the process, taking an average of 19 minutes. Then, the mean travel time after the assessment may vary relatively depending on the final destination of the patient. Thereby, travel times to B take almost double the time (19 minutes) of transferring patients to C (10 minutes). On the other hand, ambulances finishing in A are the ones that take more time on average to reach the event (15 minutes).

Table 1 also reports the coefficient of variation as the ratio of the standard deviation to the mean of the time analysed. This coefficient indicates the underlying variability in the system and it is inversely related to how well the system performs. Green (2006) argued that waiting time and service use are related positively. Once the capacity reaches a certain level of utilisation the waiting time increases exponentially. The higher the coefficient of variation, the further this capacity is from its maximum level of utilisation and it is more likely to operate inefficiently. Hence, under the same capacity, a higher coefficient of variation entails longer delays for receiving the service under the same level of utilisation. In our case study B is the centre that shows most variability in its demand considering all the phases of the pathway presenting an average coefficient of variation of 0.77 against 0.67 for A and 0.63 for C. This
may suggest that centre B may find it more difficult to adjust its capacity as demand varies and consequently there would be greater delays for a given level of resource use.

4. **System’s performance associated with various service configurations**

Table 2 shows the performance of the ambulance by means of the waiting time ($W$) and their bootstrapped confidence intervals under different scenarios. We compare the system performance baseline *All Centres* with scenarios without centres $A$, $B$ and $C$ respectively. Likewise, these comparisons take into account the 3 day-shifts defined above (columns 1, 7 and 13). In addition, we incorporate two arbitrary modifications of the baseline server capacity in order to illustrate how the system behaves under extreme plausible values. Particularly, we examine the system performance with 50% less ambulances per unit (5 servers) and also with a capacity 50 times greater (500 servers).

We control the heterogeneity of different journeys by segmenting the initial sample according to three different journey lengths of time. We thus define $G1$ for those journeys (1,118) that last less than 20 minutes, $G2$ for journeys (18,704) that last between 20 and 60 minutes and finally $G3$ for journeys (3,801) longer than 60 minutes. In order to calculate the performance variable of the system we bootstrap the explanatory parameters of the model ($\mu$ and $\lambda$) 2000 times in order to obtain their mean values within 95% confidence intervals. Then we execute the estimations of the waiting time for each of the configurations defined above (*Scenario 1* and *Scenario 2*).

**4.1 Waiting time when altering the number of ambulances**

Regardless of the scenario proposed and the type of journey, altering the number of ambulance servers does not affect notably the system performance. Considering the baseline scenario where there are 3 centres, changes in the number of ambulances do not lead to changes in the time required to provide the service. In other words, there are no economies or diseconomies of scale given the demand for ambulance services over the range of ambulances considered.

**4.2 Waiting time when altering the available centres**

The system’s performance shows greater variations in the completion time of the ambulance service when altering the number of centres than when modifying the number of ambulances. We compare the time difference in the time for completing the service between a hypothetical scenario disregarding one centre and the reference scenario with the three centres $A$, $B$ and $C$. 
This time difference may be interpreted as the cost of redesign associated with the extra time to deliver the service derived from longer journeys.

A theoretical centralisation of emergency services without centres A, or B or C would generally increase the average time required to complete the journey, although the impact varies according to which centre is removed. Greater differences in time associated with a scenario without centre A would be reflected in G1 journeys. Particularly, the average duration of these journeys would increase between 5 and 7 minutes depending on the shift. Similarly, a scenario where centre B is removed, would lead to the highest increases on the average time for G2 and G3 journeys. This would be especially the case for G3 journeys where journey time would increase by 10 minutes on average compared to the reference scenario where three centres are in use. Finally, a situation where centre C is removed would lead to an increase of 8 minutes on average for G1. Whilst G3 journey lengths would remain on virtually unchanged.

Generally, these results suggest that G2 journeys suggest a lower relative cost of re-design compared with G3 and G1 journeys. G1 journeys have a higher relative increase in the average journey time for scenarios without centres A or C. Conversely, G3 journeys are most affected when centre B is not considered. These data suggest that more resources are allocated to tackle these more critical cases. An alternative decision rule that might be explored within an economic evaluation would consist of sending those patients who can bear longer time durations to treatment without ill-effect to centres that are further away so that more critical patients can be sent to a closer centre equipped to deal with the clinical need.
<table>
<thead>
<tr>
<th>Structure/Shift</th>
<th>Number of servers</th>
<th>G1 Journeys (n=1118)</th>
<th>G2 Journeys (n=18704)</th>
<th>G3 Journeys (n=3801)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>λ( all shifts)</td>
<td>μ( all shifts)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All centres</td>
<td>0.0021 (0.0019,0.0023)</td>
<td>0.0446 (0.0409,0.0423)</td>
<td>0.0408 (0.0399,0.0418)</td>
<td>0.0250 (0.0240,0.0252)</td>
</tr>
<tr>
<td>W (min)</td>
<td>17.81 (17.65;17.95)</td>
<td>17.81 (17.65;17.95)</td>
<td>38.66 (38.51;38.80)</td>
<td>38.81 (38.66;38.96)</td>
</tr>
<tr>
<td>Night</td>
<td>17.86 (17.57;18.13)</td>
<td>17.86 (17.57;18.14)</td>
<td>37.71 (37.42;38.03)</td>
<td>37.73 (37.41;38.04)</td>
</tr>
<tr>
<td>Morning</td>
<td>17.68 (17.35;17.99)</td>
<td>17.68 (17.36;18.01)</td>
<td>39.76 (39.54;39.98)</td>
<td>40.16 (39.91;40.39)</td>
</tr>
<tr>
<td>Evening</td>
<td>17.84 (17.62;18.06)</td>
<td>17.84 (17.63;18.04)</td>
<td>37.99 (37.76;38.21)</td>
<td>37.99 (37.76;38.22)</td>
</tr>
<tr>
<td>∆W (min)</td>
<td>5.18 (4.39;6.01)</td>
<td>5.18 (4.36;6.99)</td>
<td>5.61 (5.35;5.86)</td>
<td>5.61 (5.35;5.85)</td>
</tr>
<tr>
<td>No A</td>
<td>2.17 (1.94;2.41)</td>
<td>2.17 (1.98;2.47)</td>
<td>2.08 (1.82;2.63)</td>
<td>2.08 (1.82;2.65)</td>
</tr>
<tr>
<td>Night</td>
<td>2.08 (1.82;2.65)</td>
<td>2.09 (1.82;2.65)</td>
<td>2.41 (2.04;2.78)</td>
<td>2.41 (2.04;2.79)</td>
</tr>
<tr>
<td>Morning</td>
<td>2.41 (2.04;2.79)</td>
<td>2.55 (2.16;2.93)</td>
<td>4.57 (3.45;5.77)</td>
<td>4.57 (3.42;5.79)</td>
</tr>
<tr>
<td>Evening</td>
<td>2.17 (1.94;2.41)</td>
<td>2.22 (1.98;2.47)</td>
<td>3.94 (3.04;4.84)</td>
<td>3.94 (3.04;4.85)</td>
</tr>
<tr>
<td>No B</td>
<td>6.68 (6.11;7.23)</td>
<td>6.68 (6.09;7.23)</td>
<td>5.97 (5.39;6.54)</td>
<td>5.97 (5.39;6.54)</td>
</tr>
<tr>
<td>Night</td>
<td>5.97 (5.39;6.54)</td>
<td>5.99 (5.41;6.56)</td>
<td>5.34 (4.94;5.71)</td>
<td>5.34 (4.95;5.71)</td>
</tr>
<tr>
<td>Morning</td>
<td>5.34 (4.94;5.71)</td>
<td>5.71 (5.28;6.13)</td>
<td>7.14 (6.27;8.05)</td>
<td>7.14 (6.29;7.99)</td>
</tr>
<tr>
<td>Evening</td>
<td>5.71 (5.31;6.10)</td>
<td>5.99 (5.54;6.46)</td>
<td>8.35 (7.23;8.28)</td>
<td>8.35 (7.23;8.32)</td>
</tr>
<tr>
<td>No C</td>
<td>3.00 (2.84;3.16)</td>
<td>3.00 (2.84;3.16)</td>
<td>5.62 (5.17;6.07)</td>
<td>5.62 (5.17;6.07)</td>
</tr>
<tr>
<td>Night</td>
<td>5.62 (5.17;6.07)</td>
<td>5.64 (5.21;6.08)</td>
<td>5.33 (4.99;5.64)</td>
<td>5.33 (4.99;5.64)</td>
</tr>
<tr>
<td>Morning</td>
<td>5.33 (4.99;5.64)</td>
<td>5.70 (5.34;6.06)</td>
<td>7.50 (6.74;8.24)</td>
<td>7.50 (6.77;8.22)</td>
</tr>
<tr>
<td>Evening</td>
<td>5.70 (5.34;6.06)</td>
<td>5.89 (5.54;6.46)</td>
<td>8.35 (7.23;8.28)</td>
<td>8.35 (7.23;8.32)</td>
</tr>
</tbody>
</table>

Table 2. Performance of the system
5. Discussion

The inclusion of explicit capacity constraints is not normally considered in the majority of economic evaluation models. Yet, neglecting the capacities of the system and the effects derived from them may lead to incorrect conclusions about cost-effectiveness. Our case study has illustrated that alternative ways of organising emergency care may affect the time taken to attend patients’ needs and, by implication effectiveness and cost-effectiveness of treatment.

Changes to the time taken to deliver care may not just affect health outcomes but also influence the organisation of a health service and therefore the type of health resources that are required. Thus, centres that receive a higher proportion of patients whose care has been delayed as a result of extra travel times, may require more skilled staff and need additional specialised resources to cope with potential worse health conditions and poorer prognosis at arrival. To mitigate this, systems may need to optimise the economies of scale and scope involved in their processes in order to reduce the time to treatment of patients.

Our results show how different distributions of resources would affect the time spent for completing the initial part of the emergency care; a key determinant of a good performance of the whole service. Particularly, we have studied how different hypothetical concentrations of services associated with alternative organisational structures of A&E services, could affect the performance of the system.

The modelling approach adopted may contribute to improving the quality of economic evaluations in emergency care and other similar areas where unscheduled queues are created. Further work could examine the determinants that may influence the waiting time within the time to treatment window as well as the design of polices to overcome possible bottle necks. The analysis reported here has reflected the importance of time of day when the need for care occurs on system performance. Finally, the analysis may help to inform and design future evaluation but also future interventions. Specifically, queueing may be employed to illuminate the impact of changes in the concentration of a service. This would allow alternative scenarios to be considered prior to any empirical service change is attempted.

With respect to the case study used here and considering economic evaluation more generally, the model required would need to be more comprehensive. It should capture more detailed information about the functioning of emergency services and incorporate data on the health outcomes and resource uses associated with the health centres studied. This would allow a more
rigorous evaluation of the consequences of a hypothetical consolidation of emergency services. Such an evaluation would incorporate a network of queues including not only the waiting time associated with the ambulance service but also the queues formed elsewhere in the care pathway such as those that may occur in the A&E waiting rooms.

Considering the characteristics of the dataset and the computer demands, the implementation of a queue model has been shown tractable and not as costly in terms of implementation and information requirements as other simulation approaches. Indeed, as noted by Green (2006) queuing models are not excessively data demanding and allow for the calculation of performance measures through simple formulae. Compared with state transition models, which are still commonly used in economic evaluation, another major advantage is that it is more readily possible to incorporate capacity constraints as to reflect the underlying dynamics within the system. Nevertheless, the inclusion of capacity constraints and the dynamics of the system are characteristics that may be incorporated alongside alternative approaches such as DES which constitute a flexible approach that entails fewer assumptions about the arrival processes and the service time (Kolker, 2010). DES are suitable when considering a heterogeneous population, where there are interactions between individuals and also time varying effects. However, despite these advantages and being recommended for use in economic evaluations Caro et al. (2010), the suitability of models based on DES depends notably on the management of the input data used to populate them and the consequent time devoted to perform the simulation. Because DES are computationally demanding such models may lack transparency to the lay reader (Karnon and Afzali, 2014).

Our model is built under the assumption that the arrival process follows a homogenous Poisson distribution. The use of this distribution may result in an overestimation of waiting times (Green 2006). Taking into account this and the fact that our data reflects seasonal patterns, further research might explore other distributions and incorporate methods based on structural time series analysis or ARIMA models to address this seasonality as argued in (McClean et al., 2011).

Likewise the capacities included in the model are simplified and stylised. We are assuming that health centres only receive patients arriving by ambulance. However, patients may arrive by other means and these may compete with patients that arrive by ambulance for the same resources in the health centres. We are also assuming that the service begins without previous activity in the system. In addition, we have not considered prioritising care for severe cases nor
pre-emptive cases. Including these alternatives would imply blocking events within the model which might alter the results associated with the different scenarios proposed. Pre-emptive queues are difficult to consider in a framework such as the one presented in this paper though. In order to preserve safety is unlikely that ambulances leave patients in the event to attend other priorities. Yet, pre-emptive queues could be used for designing the protocols to follow during the journey to the health centres in order to handle with blocks in cases of heavy traffic. Despite not including these complexities, the discussion around the value of this approach within an economic evaluation is still valid. Furthermore, taking an economic evaluation perspective, even at its simplest the queuing approach provides no less information than a state transition model where resource constraints and the possibility of blocking are typically disregarded. This approach enables the introduction of explicit constraints at a reasonable computation cost.

6. Conclusions

This paper has presented the potential of queuing modelling for use within economic evaluations considering the organisation of a specific health service. In addition to the capabilities of this approach to simulate different scenarios, complementary economic analyses should explore further consequences associated with the different organisational structures depicted by each of the models presented. These studies would include research on the possible schemes of care services that could be provided in the centres as well as the optimal size of the facilities. This would contribute to better understand how services are delivered and therefore address the future challenges faced by the National Health System.
References


Huang Q, Thind A, Dreyer JF, Zaric GS. The impact of delays to admission from the emergency department on inpatient outcomes. BMC emergency medicine 2010;10; 16.


Little JDC. A proof for the queuing formula: L= λ W. Operations research 1961;9; 383-387.


Wilde ET. Do emergency medical system response times matter for health outcomes? Health economics 2013;22; 790-806.