The development and application of an analytical healthcare model for understanding and improving hospital performance

Daniel Suen,

Submitted for the degree of Doctor of Philosophy

at Lancaster University.

September 2015

Abstract

Healthcare systems are tasked with balancing a variety of conflicting priorities such as increasing patient demand, minimising waiting times and a limited budget. With an ageing population the pressure on hospitals to maintain the quality of patient care is only going to rise as demand increases. Improving the management of these systems is important for avoiding potential dangers such as patient overcrowding and bed blockages which can result in or exacerbate problems such as increased risk to patient care, excessive patient lengths of stay, and staff burnout.

This thesis develops and applies an analytical model, which uses queueing theory to approximate healthcare systems in order to provide a means of analysing these systems mathematically. In particular we consolidate, simplify and extend the theory underpinning both single node and networks of infinite server queues, using ideas and concepts developed in Gallivan and Utley (2005) and Massey and Whitt (1993) in order to derive explicit and easy to use formulae for the mean and variance of bed demand.

We demonstrate the use of the analytical model by using the model outputs to produce model-based performance indicators to measure hospital performance and hence identify hospitals deserving further investigation. A difficulty in evaluating hospitals is determining how to measure their performance and produce fair and meaningful results while accounting for factors beyond their control, for example hospital size impacts on the relative variability of patient demand and needs to be incorporated into any analysis. We analyse the elective and emergency work of 30 hospitals using model-based performance indicators as a point of comparison

for the observed results, allowing for hospital size. For the emergency work we focused on a single length of stay distribution but a key difference arose in the elective case, where we incorporated day-of-week dependent patient lengths of stay. In cases where day-of-week dependent length of stay data is not available, we also devise and evaluate a statistical approach for model calibration.

Acknowledgements

I have a received a great deal of help and support throughout my time working on this thesis. Firstly a huge thank you to my family, in particular my parents and sister Carmen - and of course our cat Dipple. I could not have reached this point without all of your love and encouragement.

My supervisory team Matt Sperrin and Dave Worthington. I am incredibly grateful for the support and guidance you both have shown me. In particular I would like to thank Dave for his never ending patience and kindness throughout the last four years. And thank you to the STOR-i department for providing me with the opportunity to do this research.

To Jak and Erin, my office mates. Thank you for always sharing my fears and worries. And most importantly for introducing me to "The Settlers of Catan". Without your humour and friendship this work could not have been possible.

Declaration

I declare that this thesis is my own work except where noted otherwise and has not been submitted for the award of a higher degree elsewhere.

Daniel Suen

Contents

Chapter 1

Introduction

1.1 What is operational research?

Operational research (OR) has been defined by the Institute for Operations Research and Management Sciences (INFORMS) as "the discipline of applying advanced analytical methods to help make better decisions". In other words OR is the broad area of using scientific methods to aid and improve decision making in some organisation.

The definition of OR is not confined to the method or tool used to solve some organisational problem but extends to the entire problem solving process. OR is a systematic approach to problem solving and this includes identifying and defining the problem, constructing a model and selecting the best or most appropriate solution. Identifying and defining the problem often involves meetings and discussions with industry experts and managers experiencing the issue to better understand the problem at hand.

1.2 Why is OR important?

OR is important for organisations to provide products and services as efficiently and effectively as possible. This can be achieved using scientific methods and techniques to support and inform decision making. This requires careful planning and

analysis and can employ statistics, computing, optimisation, and mathematical modelling to assist in solving an array of business and organisational problems.

With the growth of technology and the global economy effective and informed decision making has become all the more important for modern day organisations. By using sophisticated statistical and mathematical models, OR users are able to simplify increasingly complex problems to assess and evaluate all available options. As a result management decisions are made on more reliable and complete information.

Underpinning OR methods is an objective and quantitative basis for decision making and therefore solutions found are an improvement on one based purely on the (subjective) opinion of experts in that it will lead to better and more consistent decision making. That is not to say expert opinion is precluded from the decision making process but is rather complementary to it. OR should be viewed as an aid to good decision making.

1.3 OR modelling

Applying OR to a problem involves some sort of model formulation. This is because it is typically easier to analyse a simplified model of some real world process, drawing conclusions and solutions from the model which can be presented to the decision maker. Models are not designed, nor expected to, accurately represent every feature within a process, rather they are an approximation to the real world process that are able to capture the salient features. Interpreting and understanding any results should always be done while conscious of the assumptions made in the model.

1.3.1 Computer simulation

Computer simulation models are models whereby the system is abstracted into a computer program. They are used to test and analyse the behaviour of systems

under different operating conditions cheaply and without risk to the real world operations. Often computer simulation software provides nice and easy to understand graphics and animations which help both modellers and decision makers visualise the system being simulated. Models which are better understood are naturally more trusted and potential solutions are more likely to be implemented by managers.

The time it takes to run a simulation can vary from a few seconds, to days or even longer. The inputs of simulation models include decision variables and are under control of the user. This allows users to experiment with different combinations of input variables, letting them observe the resulting behaviour of the simulated system under different scenarios. Statistics are gathered on various measures of performance at the end of a simulation run which can then be analysed.

Simulation models can capture complicated and important network effects without necessarily understanding all of the workings of the system. This facet of simulation modelling can be described as a black box. There is a drawback to this black box approach, since the user does not need to understand all of the interactions and processes within the system this can mask logical errors in the model (Fraser, 2010).

Another drawback to simulation models is that while they capture important system features to simulate behaviour they do not produce preferred solutions to improving the system, rather they evaluate a system and produce outputs under certain conditions. This means simulation models do not produce preferred solutions to improving the system, rather they evaluate a system and produce outputs under certain operating conditions specified by the modeller. This process is often referred to as “what if” modelling

1.3.2 Queueing models

Queueing theory is the study of waiting lines or queues based on probability theory and mathematics and is often used when making decisions about the resources

required to provide some service. Queueing models consist of three major components:

- how customers arrive,
- how customers are serviced, and
- the number of servers available.

Often queueing models are described using Kendall's notation in the form $A/B/c$ where A describes the arrival process, B the service time distribution and c the number of servers (Kendall, 1953). This notation was designed for queues in continuous time so care must be taken when using this in relation to queues in discrete time

Unlike computer simulation, analytical queueing models offer a different approach which requires a deeper understanding of the system. A benefit of queueing models is that queues can be described using simple and elegant mathematical equations which can be used to analyse and understand the queue behaviour. Analytical queueing models can be used to estimate such performance metrics as customer throughput, mean queue lengths and mean sojourn time. Another consequence of the equations derived from queueing models is that they can be constructed such that they lend themselves to optimisation programs which can used to determine optimal operational plans. However as more complexity is added in the system, closed-form formulae become less and less tractable.

1.3.3 Optimisation models

Optimisation models, sometimes known as *mathematical programming models*, allow users to find the best or "optimal" solution when many feasible options are available. In this type of model there are three main elements:

- decision variables,
- constraints, and

- an objective function.

Decision variables are features of the system which are under the control of the user. Optimisation models will search different combinations of decision variables until a specific set of values is found such that the objective function (sometimes known as a loss or cost function) is maximised (or minimised). Constraints are limits or bounds on the range of values which the decision variables can take. Typically constraints represent some sort of restriction on for example staffing hours, or requirement such as contractual obligations. The objective function is mathematically stated measure of performance such as cost, profit, time, etc., which is a function of the decision variables.

Often managers have more than one objective e.g. maximise profit while minimising changes in workforce levels. In such cases modellers can either focus on a single objective and place the alternative in the constraints, agreeing upon a specified maximum or minimum value which it can take or use a technique designed specifically for multiple objectives. One example would be to used some weighted combination of the two objectives, however this would require defining the relative importance (weights) of each objective which is not necessarily straightforward.

1.4 Statistical methods

Statistical modelling is similar to what we have described as OR modelling in some respects. It involves simplifying a real world process using a set of mathematical equations. The variables in the equations do not necessarily have specified values but may be characterised using probability distributions. We present some statistical methods which are particularly useful in a healthcare context.

1.4.1 Forecasting

Forecasting is the science of making predictions on the future behaviour of a system based on past and present data and analysis of trends. It is a planning tool to help

management cope with the uncertainty of the future. Estimates or predictions are made based on a variety of statistical techniques such as exponential smoothing, moving averages and regression analysis to name a few.

Forecasting is used in many industries for example, forecasting the call volume at a call centre can help schedule staff appropriately. Our ability to forecast some quantity accurately depends on a variety of factors including our understanding of contributing factors and the amount and quality of data available. With this in mind the level of uncertainty attached to forecasting is important and modellers should indicate the degree of uncertainty.

1.4.2 Performance indicators

Performance indicators are a type of performance measurement. They provide information to the user about the behaviour of the system of interest. Selecting good performance indicators requires a good understanding of the system and what is important to the decision makers. Unlike say optimisation techniques which derive a specific optimal operational plan, performance indicators can be used to provide a profile of the system performance to assess how well or poorly the system is doing and identify potential areas for improvement. Units, for example hospitals, scoring well according to performance indicators are sometimes used to flag up examples of good practice.

There are some dangers with using performance indicators. If not selected carefully such that a performance indicator or set of indicators fairly reflects a system then people may focus on improving performance indicators and not necessarily the underlying objectives. With this in mind it is important for modellers to select performance indicators carefully and with an understanding of the system.

1.5 Statistics and OR in healthcare

In a healthcare context statistics and OR are of particular importance, and are often used together. For many years hospitals have been facing steadily rising patient numbers and a shrinking budget. Rising pressure to maintain quality of care while coping with a limited budget motivates the need to improve the management of healthcare systems. Royston (2009) provides a nice non-systematic review of the past, present and future contributions of OR in health in the UK. Going back as far as 1948 when the National Health Service (NHS) began in the UK, Royston broadly examines the growth and contributions of OR in healthcare - illustrating this with an example of how OR work has been published every decade since the 1950s on waiting times in accident and emergency departments. But Royston (2009) also indicates the difficulties faced by health OR in the UK, citing instances where the use of OR has not been more readily applied or had as big an impact as it could possible warrant e.g. "Compared with many other organisations, hospitals have been slow in adopting operational research as a means to improve their performance" (Buhaug, 2002).

Proudlove et al. (2007) raise another important issue in the challenges of applying OR modelling effectively in the NHS. The complexity of models can often be unnecessary and difficult for problem owners to understand - hindering the provision of useful insight and guidance to problems. Implementation of OR techniques is a real issue in healthcare but is often not given the attention required. There is no easy solution to this however Royston (2009) identifies an important chain of requirements which OR techniques need to be when it comes to successfully implementing OR in healthcare (and other areas): "Available, visible, relevant, affordable, comprehensible, convincing, practical and timely" (Royston, 2009). But to achieve this, it requires active involvement from both the health OR community and the problem owners - healthcare managers, senior clinicians/decision makers, etc.

Bennett et al. (2012) recognise that the role of OR analysts in healthcare does not lie only in using relatively-advanced analytical methods but also in problem formulation, understanding and making sense of issues, in a manner which allows the application of relatively simple models. Bennett et al. (2012) expands on this idea of OR analysts having an important role in synthesising problems with the use of some examples, including the organisation of primary care within the English NHS in the form of of General Practitioners (GPs). GP practices are independent contractors which provide a service, therefore there are regular negotiations at a national level around the contract. The Primary Medical Care Analytical Team support the DH in these negotiations, needing to quickly understand, model, clarify and advise on the potential impacts of any new ideas or proposals put forward. Bennett et al. (2012) identifies the role OR analysts play in "establishing understanding, visibility and credibility for analysis" supporting some of the thoughts in Royston (2009).

The use of OR can be successful in improving healthcare systems. For example Bowers et al. (2012) develop a decision support tool for the Patient Transport Service. This component of healthcare systems is important, especially in Scotland which has a significant remote/rural population. Bowers et al. develop a simple tool for solving this type of vehicle routing problem but compare the results with more sophisticated techniques. Interestingly Bowers et al. (2012) state that some of the attractive qualities of their simpler model were its pragmatism, transparency, and being able to deliver the solution in a timely manner. These qualities are identified in Royston (2009) for the effective application of OR techniques in healthcare. This model was successfully implemented and is used by the Scottish Ambulance Staff.

Understanding patient flows in healthcare systems is an important tool when trying to improve hospital efficiency and, among other things, reduce patient waiting times. A better insight into hospital performance help decision makers improve the management of hospital resources (e.g. hospital beds, staff) and avoid patient

blockages, where a build up for one type of resource can have knock on effects on the rest of the system.

There are many ways in which statistical and OR methods are being used to improve healthcare systems. For example:

Demand forecasting: Due to the continuing growth in demand for healthcare services accurately predicting demand and cost for the future is important. Efficient healthcare systems aim to match resources to demand for services over time. Jalalpour et al. (2015) indicates the value in accurately anticipating demand for services in order to match capacity, staffing, and supplies to be as efficient as possible. In healthcare systems forecasting demand often relates to hospital bed requirements (Farmer and Emami (1990), Green (2002), Littig and Isken (2007)) where there is concern over demand exceeding capacity can cause patient blockages and backlogs or alternatively the cost of having empty beds as a waste of scarce hospital resources.

Staff scheduling problems: Scheduling problems often deal with attributing the appropriate number of resources (staff, beds, etc.) to patient demand. In the case of staff scheduling optimisation models have been used to construct optimal schedules for different types of clinical staff such as physicians (Bruni and Detti, 2014) or interns and residents (Bard et al., 2014). Staff scheduling problems are important because many services need to be provided on a continuous basis, twenty-four hours a day, seven days a week. Furthermore these types of problems are complicated by training requirements and departmental rules can make finding optimal scheduling plans difficult.

Reducing delays and waiting times: Timeliness is a major factor in patient satisfaction and quality of care. But patient delays remain a prevalent issue for healthcare systems which can have adverse clinical consequences and higher costs. For example in an A&E department problems with available capacity relative to demand is exacerbated by high variability in demand over time. OR methods such as queueing analysis has been used to produce more effective staffing levels

to accommodate fluctuations in weekly patient demand and as a result reduce patient waiting times (Green (2008), Green et al. (2001)).

Improving patient flows: Discrete-event simulation modelling can be used to model the impact of changing patient flows in healthcare systems. Konrad et al. (2013) address problems with waiting times and congestion in an emergency department and explores an alternative patient flow process, a split-flow concept. The split-flow concept involves dividing patient flow according to acuity and enabling parallel processing. In this example discrete-event simulation is used to compare the split-flow approach with other types of emergency department triage.

Modelling patient flows is directly related to patient demand. By modelling patient flows one can estimate the patient demand placed on the system and therefore the resources required. For example hospital beds are a critical resource, frequently in demand and a lack of available beds can impede the flow of patients within healthcare systems. High surgical bed occupancy levels can cause heightened staff stress, increased surgical cancellations, and increased surgical waiting times (Chow et al., 2011). Addressing the resource demand requirements can improve patient flows by identifying and reducing pressure on bottleneck points in the system.

1.6 Infinite server models

Queueing models are used to approximate real queueing systems so that the behaviour of the system can be analysed mathematically and are often primarily focused on congestion problems associated with limited resources such as waiting times and blockages. These models are able to accommodate different types of arrivals and service times.

Any model characterising patient flows within a healthcare system must incorporate a variety of complexities such as:

- Patients residing in healthcare systems are often a mixture of planned and

unplanned work,

- the stochastic and non-stationary nature of arrival and service times, and
- the network nature of hospital systems.

Including these features is crucial for developing accurate and useful models but due to the variety and types of complexities found in healthcare systems constructing a suitable and robust model is far from straightforward.

Infinite server models are types of queueing models which assume there are an unlimited number of servers. When modelling healthcare systems the infinite server assumption means all bookings and arrivals are honoured i.e. no patient is ever turned away nor do they share resources. Consequently patients in the system are independent which makes including multiple patient types relatively straightforward by treating them separately.

Infinite server models can produce closed-form formulae for the mean and variance of the number of patients in the system, and sometimes the full probability distribution. Whereas many capacitated models are limited in that tractable results are not so readily available, the infinite server assumption does not have the same limitation while still incorporating different types of patient arrivals and lengths of stay.

We should note the mismatch of the $M/M/\infty$ notation and the arrival/length of stay processes in models which operate in discrete time. In a classical queueing sense, Poisson arrivals or $M/M/\infty$ models denote a negative exponential distribution for inter-arrival times. This is not the case in a discrete time model where "Poisson arrivals" has a different meaning, implying that the number of arrivals at a particular instant is Poisson distributed. One could argue a notation $A_D/B_D/\infty$ where A_D denotes the distribution of number of arrivals at a discrete time point, and B_D denotes the discrete distribution of service time could be a discrete equivalent for the queueing notation.

Gallivan and Utley (2005) provide an example of a single node infinite server

model being used to characterise patient flow. In this case the model incorporates both elective and emergency patient types. The elective arrivals are characterised using the Binomial distribution to account for the random non-arrivals of booked patients. Otherwise we consider the elective arrivals as deterministic. Emergency arrivals are characterised using the Poisson distribution. Each patient type has its own length of stay distribution.

Utley et al. (2009) expand on the infinite server model and derive results for the mean and variance of demand in a multi-state flow system using a time-varying general arrival distribution and general patient length of stay distribution. The case presented in Utley et al. (2009) assumes the multi-state flow system is a rooted directed tree - a system where patients have a single point of entry and each other state is accessible by exactly one directed path. Patients move between "states", which may represent physical locations in a hospital, and the time spent in each state has some variability. In this case Utley et al. are able to derive formulae for the mean and variance of demand at some state i, T time units into the modelled period.

This demonstrates how suitable the infinite server model is by capturing the main factors of patient demand while still producing closed-form formulae for the mean and variance of bed demand. Helm and Van Oyen (2014) also use an infinite server model but instead addresses an entire hospital with multiple wards. This represents an extension of the single node system whereby multiple infinite server queues are connected. Conveniently infinite server theory extends to networks of interconnected queues such that closed-form formulae can still be found for the mean demand at individual nodes. This represents a different extension to that found in Utley et al. (2009) which focused on *rooted directed tree* systems. Helm and Van Oyen's approach allow patients to move between any given state to any other state to represent for example transfers between wards within the hospital as a result of a change in the patient's condition. Helm and Van Oyen (2014) use deterministic arrivals to capture the arrival process of elective (scheduled) patients.

1.7 Research aims

In this thesis we are interested in developing infinite server models for characterising patient flows, using such models to better understand the performance of healthcare systems in terms of patient demand and how this information can be used to improve and inform decision making within these systems. The overall aim of this research is to investigate the scope for the greater use of infinite server models to aid the management of patient demand.

Infinite server models can be used to characterise patient flows in healthcare systems to estimate resource requirements. The motivation for using infinite server models to characterise patient flow is that they are able to incorporate the main factors which drive patient demand while still producing closed-form formulae for the mean and variance of demand.

In order to achieve this overall aim we first consider existing theory and applications of infinite server models, identifying their successes and limitations. We then go on to consolidate and extend the existing theory before developing new types of application, which exploit existing hospital data to underpin improvement in healthcare systems.

1.8 Structure of the thesis

The remainder of this thesis begins with a literature review. Chapter 2 provides a summary of the existing theory on infinite server queues and applications of the theory in healthcare settings. We highlight some optimisation models used alongside infinite server models as a means of improving healthcare decision making. This also motivates our interest in performance indicators and how they can potentially be used in improving decision making. This chapter helps identify the limitations of the current uses of infinite server models in healthcare and suggests ways of improving the use of such models.

In Chapter 3 we extend the existing theory of single node infinite server mod-

els, providing results in the form of expressions for the mean and variance of bed demand while allowing a general arrival distribution, whereas existing theory assumes either Poisson or Binomial arrivals. This increases the scope and flexibility of the infinite server model in characterising different types of patient arrivals. Furthermore we consider Massey and Whitt (1993)'s thorough but difficult to understand methodologies when deriving results for networks of infinite servers. In Chapter 3 we offer a different and much simpler approach to computing the mean demand of individual nodes in networks of infinite server queues and also show our result is equivalent to that in Massey and Whitt (1993).

In Chapter 4 we use the infinite server model to compare the flows of emergency work of 30 hospitals of varying size from around England. We first demonstrate the value of the infinite server model to compare the performance of the 30 hospitals. We then introduce so called "model-based performance indicators", statistics using our modelled results which provide insight and understanding about the relative performance of the hospitals. This chapter offers a new approach to using infinite server models for improving healthcare systems in the form of model-based performance indicators.

Chapter 5 tackles the counterpart elective work of the same 30 hospitals seen in Chapter 4. Elective patients present a more challenging modelling problem than the emergency work, requiring day-of-week dependent length of stay distributions, and analysis of weekday demand only. We again construct model-based performance indicators, this time for elective and total hospital work. We discuss the limitations of studying each patient type individually and the importance of considering the total demand at each hospital when measuring performance.

Chapter 6 addresses an important statistical issue which arose when modelling the elective work in Chapter 5. In particular the parameters of the day-of-week dependent length of stay distributions had to be derived from data sets which only provided daily admissions and daily occupancy levels. In this chapter we investigate the main factors affecting the parameter estimation and provide a statistical

analysis into the accuracy of the calibration process.

Finally Chapter 7 summarises the main results of the thesis and discusses the findings on using infinite server models for informing and improving the decision making in healthcare systems. We also provide recommendations for future research.

Chapter 2

Literature Review

2.1 Introduction

Healthcare decision makers are under pressure to manage hospitals and other healthcare systems while balancing limited financial resources with growing patient numbers (Harper, 2002). This is no simple task when one considers the variety of conflicting objectives faced by decision makers such as patient waiting times (Mayhew and Smith, 2008), patient overcrowding (Holm et al., 2013), and limited staffing resources to name just a few.

Modelling healthcare systems is important for understanding how they work and identifying areas for improvement. Queueing models are able to capture the stochastic and time dependent nature of patient arrivals and lengths of stay, and hence they are useful in characterising patient flows within healthcare systems. We are particularly interested in *infinite server models* and how they can be used to help inform decision makers and improve hospital management.

Literature on infinite server models were found from searching the references from key papers (Gallivan and Utley (2005) and Helm and Van Oyen (2014)), leading to other important papers such as Massey and Whitt (1993). We also searched for other papers published by the same authors providing other relevant literature such as Utley et al. (2009).

Infinite server models are useful for a variety of reasons; in comparison to

simulation models they are simpler, cost less, easier to implement and provide more generic results (Fomundam and Herrmann (2007), De Bruin et al. (2007)). Although we should be aware that simulation models offer a more user friendly graphical interface and are more easily understood by doctors and managers - and therefore can be more convincing. Infinite server models also provide analytically tractable results. There is a need to plan and manage hospital resources with particular emphasis on inpatient bed management (Gorunescu et al. (2002)), infinite server models offer a means of computing the mean and variance of bed demand which can be used, for example, in optimisations to find optimal elective admission plans (Gallivan and Utley (2005), Helm and Van Oyen (2014)).

Infinite server models use the modelling concept of "unfettered demand" (Utley and Worthington, 2012), the potential demand of the system assuming all arrivals are honoured and never turned away. This brings us to perhaps the most obvious concern with infinite server models, the fact that an unlimited supply of servers is often unrealistic and impractical - hence why use such a model? Even with this unlimited capacity assumption, infinite server models still offer insights into capacitated systems. In some cases assuming sufficient capacity is desirable because patient arrivals cannot be turned away, for example an emergency department would make every effort to accommodate patients which are in critical condition, where turning them away would potentially pose a serious risk to their health. In contrast capacitated queueing models do not have the same convenient results as their infinite server counterparts, making them more difficult to analyse and encouraging the use of infinite server models to approximate capacitated systems.

We review the infinite server theory and healthcare applications in order to identify possible areas for further work. The following Section 2.2 focuses on the theory of single node infinite server queues. We provide results for both homogeneous and nonhomogeneous infinite server models, including the distribution of the number in the queue and the distribution of the departure process. Section 2.3 reviews applications of single node models, highlighting a range of papers which

use single node queueing theory to characterise patient flow in different types of healthcare systems. We then introduce networks of infinite server queues in Section 2.4, summarising results for networks with time dependent Poisson arrival processes and highlighting the important concepts of decomposition of networks and the departure process of individual nodes within the system. Section 2.5 identifies two different healthcare applications of the infinite server network theory; Helm and Van Oyen (2014) is an example of infinite server theory being used with optimisation and Izady and Worthington (2011) an example of using the mean bed demand in networks to compute staffing levels in a hospital. We describe the way in which these papers use the relatively complex network theory to characterise their healthcare systems and how they make use of the underlying assumption of infinite servers. In Section 2.6 we provide an in depth comparison of Gallivan and Utley (2005) and Helm and Van Oyen (2014). These two papers use the infinite server models in conjunction with optimisation procedures to derive optimal elective admission schedules. We also include in the comparison a third paper, Bekker and Koeleman (2011), which focuses on the issue of the variability of bed demand using an infinite server model and uses optimisation to find an optimal admission schedule. Section 2.7 discusses other possible uses of infinite server models in healthcare systems beyond optimisation such as performance indicators. Section 2.8 explains what performance indicators are and how they can be useful in a healthcare context. Finally in Section 2.9 we discuss the literature, summarising our findings and noting limitations in the current applications of infinite server queueing theory in healthcare systems which will help clarify the direction of the research in the remainder of the thesis.

2.2 Single node infinite server queueing theory

Well established results for the $M/G/\infty$ model have been around for many years where the M indicates a homogeneous Poisson arrival process, G indicates a general length of stay distribution and the number of servers is infinite. The single node

homogeneous arrival model forms the basis of many infinite server queueing models, in particular its nonhomogeneous counterpart, $M_t/G/\infty$ (where M_t indicates a nonhomogeneous Poisson arrival process), and the extension to networks of infinite server queues which we focus on later in this chapter. In this section we summarise the main results of the single node infinite server queue with both homogeneous and nonhomogeneous arrivals.

2.2.1 Homogeneous arrivals

In this section we review established results for the $M/G/\infty$ model (proofs of which can be found in Tijms (2003)). Early applications of the $M/G/\infty$ model can be found in a telephone traffic context (Riordan (1951), Benes (1957)) where certain key properties, such as the number of customers in the system is Poisson distributed, are derived.

The homogeneous $M/G/\infty$ queue differs from the time-dependent case in that we can analyse its steady-state behaviour. Steady state is when sufficient time has elapsed so as the current state is independent of the initial state i.e. has settled down into some kind of equilibrium position.

Consider an $M/G/\infty$ model, the queue length at time t is Poisson distributed with mean:

$$m(t) = \alpha \int_0^t G^C(u)du \tag{2.1}$$

where

- α is the arrival rate, and
- $G^C(t) = 1 - G(t)$ is the probability the patient has not left the system by time t, i.e. $G^C(.)$ is the survivor function associated with the length of stay (see Newell (1966)).

Note that since the queue length is Poisson distributed, the mean queue length (Equation (2.1)) is also the variance of the queue length at time t.

Furthermore in the steady state analysis, when we let $t \to \infty$, then as:

$$\int_0^\infty G^C(u)du = \mu.$$

Then the limiting distribution of the number of customers in the system is a Poisson distribution with mean $\alpha\mu$:

$$\lim_{t\to\infty} P(k \text{ servers are busy at time } t) = e^{-\alpha\mu}\frac{(\alpha\mu)^k}{k!},$$

for $k = 0, 1, 2, ...$ Note that this limiting distribution does not require the shape of the length of stay distribution, only the mean.

Since the customers leaving the $M/G/\infty$ queue may enter some other queue, it is useful to analyse the departure process. The number of customers leaving at time t is also a Poisson process and has rate:

$$\delta(t) = \alpha \int_0^t g(u)du \tag{2.2}$$

where $g(t)$ is the probability a patient spends time t in the system, i.e. the probability density function of the length of stay distribution. Furthermore in the steady state analysis of the $M/G/\infty$ queue, when we let $t \to \infty$, then the departure process is also a Poisson random variable with parameter:

$$\delta(t) = \alpha \tag{2.3}$$

i.e. the departure process is the same as the arrival process (see Gautam (2009)).

These are two important results, the distribution of the number of customers in an $M/G/\infty$ queue being Poisson distributed and the departure process also being a Poisson process, for which we next show the equivalent results in the nonhomogeneous case.

2.2.2 Nonhomogeneous arrivals

We now focus on an extension of the $M/G/\infty$ model, where we have a time-dependent Poisson arrival process in the $M_t/G/\infty$ queue. Everything of the $M_t/G/\infty$ model is the same as the homogeneous case except for the arrival process which is still Poisson but the parameter is time varying, $\alpha(t)$. Even with this additional feature, the main results for the number in the system and the departure process are still obtainable which we present here (see Eick et al. (1993) for proofs).

The queue length of the $M_t/G/\infty$ model at time t is Poisson distributed with mean:

$$m(t) = \int_0^t \alpha(t-u)G^C(u)du. \tag{2.4}$$

Note that this result is equivalent to that in the $M/G/\infty$ case, Equation (2.1), with the exception of the arrival rate is now a function of time.

The number of customers leaving the $M_t/G/\infty$ queue, the departure process, is a nonhomogeneous Poisson process with rate function:

$$\delta(t) = \int_0^t \alpha(t-u)g(u)du. \tag{2.5}$$

Also for each t the queue length is independent of the departure process.

2.3 Single node infinite server theory: healthcare applications

Single node infinite server models are used to characterise healthcare systems to estimate bed demand while incorporating multiple patient types, homogeneous and time dependent arrivals, and patient length of stay distributions. Infinite server queueing models are useful because they offer tractable and easy to compute results

on the number of patients in the system.

An important feature of healthcare systems is the multiple types of patient which pass through their systems. Capturing different streams of patients is important when modelling healthcare systems because they affect the performance of the system. Ignoring one or more different types of patients can be inaccurate or misleading in terms of the information and results produced on the system.

Patients may be defined by the type of illness or reason for arriving at the healthcare system. In modelling terms this typically means patients of different types are defined by their arrival pattern and length of stay distribution. For example categorising patients as either elective or emergency and defining these two groups by the differing nature of their arrivals. How one chooses to divide patients is up to the modeller. Infinite server queueing models are flexible in that incorporating additional patient types is relatively straightforward, taking advantage of the fact that patients are independent through not competing for resources.

An obvious caveat of infinite server models are that they may be considered unrealistic due to the sufficient capacity assumption, where in reality a lack of beds is a serious issue for for patient flow and overcrowding (Proudlove et al. (2003), Derlet et al. (2001)). This concept of "unfettered demand" (Utley and Worthington, 2012) i.e. patient arrivals are always honoured and never turned away, is important for healthcare managers as frequently they are interested in the potential demand of their systems rather than the demand when constraints are placed on capacity.

For example De Bruin et al. (2007) acknowledges infinite beds is not always realistic but analysing the number of beds required to accommodate all arrivals is still relevant because one of the main goals of the hospital is providing an admission guarantee for all arriving patients. In some cases turning patients away is simply not an option therefore planning for potential patient demand with no capacity constraints is sensible.

A convenient result of the $M/G/\infty$ model, where arrivals are assumed to follow a Poisson process is the number of patients in the system is also Poisson distributed. De Bruin et al. (2007) uses the distribution of the number of beds occupied to compute the probability of the number of beds required exceeds the number of beds present in each of their three units which they use to help determine bottlenecks in their system.

Single node infinite server models also have some useful results when modelling non-Poisson arrivals. Utley et al. (2003) argue the importance of predicting bed demand when trying to plan elective patient admissions in advance while trying to minimise the number of cancellations by maintaining a certain level of reserve capacity. It is this uncertainty in predicting bed demand that leads them to construct closed form expressions for the mean and variance of bed demand for elective and emergency patients which are defined by their Binomial and Poisson arrivals respectively.

Gallivan and Utley (2005) build on the results from Utley et al. (2003) for the mean and variance of bed demand. In this case they construct an optimisation incorporating the mean bed demand to determine an optimal elective admission schedule. At the optimisation stage of their modelling process they incorporate a limit on bed capacity. Bekker and Koeleman (2011) take a similar approach, using an approximate infinite server model ($G/G/\infty$) to compute the mean bed demand and incorporate this into an optimisation. They seek to produce an optimal admission schedule by reducing the variability in bed demand, minimising the difference between the mean bed demand and some target load. In both cases the idea of what is "optimal" remains subjective and fails to incorporate the variance of bed demand in the optimisation (though Gallivan and Utley (2005) do use the variance of bed demand to analyse the results from their optimal admission schedule). The use of optimisation to improve healthcare systems using infinite server models is of particular interest and both of papers (Gallivan and Utley (2005) and Bekker and Koeleman (2011)) are subject to closer comparison in Section 2.6.

2.4 Networks of infinite server queues

Networks of infinite server queues are an extension of the single node case where a network is a set of interconnected queues, with customers in the system able to move from one node to another. Massey and Whitt (1993) focuses on networks of infinite server queues with nonhomogeneous Poisson arrivals providing important concepts and results for these types of systems. In particular we are interested in the mean demand at individual nodes in a $(M_t/G/\infty)^{N_1}/M$ network where N_1 represents the number of nodes in the network and the latter M indicates Markov routing (where the queue sequence is independent of the arrival process and service times and determined by the transition probabilities).

Massey and Whitt's method for computing the mean demand of individual nodes in this type of system introduces two important ideas; decomposition of Markov route networks into a set of tandem node networks with deterministic routes and using the departure rate (of a previous node) as the arrival rate of the subsequent node. Tandem node networks are a sequence of nodes where patients move from one node to the next following some deterministic path and cannot visit any previous node (should this be the case, subsequent visits are characterised as a separate new node). This section provides the key results for networks of infinite server queues from Massey and Whitt (1993), using the important concepts of decomposition and the departure rate function, highlighting why their methodology is complex and difficult to implement.

2.4.1 Theory of $(M_t/G/\infty)^{N_1}/M$ networks

According to Massey and Whitt (1993) the queue lengths at time t for node k, $1 \leq k \leq N_1$, are independent Poisson random variables with finite means

$$m_k(t) = \int_0^t \alpha_k(u) G_k^C(t-u) du \tag{2.6}$$

where

- $\alpha_k(t)$ is the arrival rate to node k at time t, and
- $G_k^C(t) = 1 - G_k(t)$ is the probability the patient has not left node k by time t, i.e. is the survivor function of node k at time t.

Furthermore the aggregate departure process (departures to other queues as well as departures from the system entirely) from queue k is a Poisson process if and only if no customer can visit queue k more than once. We note the close relationship between the distribution of the demand of individual queues in a $(M_t/G/\infty)^{N_1}/M$ network and the distribution of an isolated $(M_t/G/\infty)$ queue (Equation (2.4)).

2.4.2 Decomposition

To derive the mean demand of queue k at time t Massey and Whitt decompose the original $(M_t/G/\infty)^{N_1}/M$ network into a set of $(M_t/G/\infty)^N/D$ networks where D indicates deterministic routing. Note that N may be greater than than N_1 as patients may visit a node more than once. Customers in the network are grouped by their path through the system. All customers which share the same (deterministic) path are grouped together, and each group of customers are modelled by an $(M_t/G/\infty)^N/D$ tandem node network. In each of these deterministic route networks customers follow a fixed path and should any node be visited more than once we characterise each subsequent visit as a separate (new) node. The arrivals to any given node are the departures from the preceding node.

It is convenient to think of the original network as a multi-class model which we break down into a set of single class deterministic route networks - where each customer class follows a deterministic path. The number of people at individual nodes in each $(M_t/G/\infty)^N/D$ network is a Poisson random variable. Therefore we need only consider a single $(M_t/G/\infty)^N/D$ network to compute the mean demand of an individual node, the desired result for the original $(M_t/G/\infty)^{N_1}/M$ network is obtained from the sum of the contributions from each $(M_t/G/\infty)^N/D$ network.

2.4.3 Departure rate function

To compute the mean demand of node k in a $(M_t/G/\infty)^N/D$ network the arrival rate function of node k is required. Massey and Whitt note that the departures of the previous node are the arrivals to the following node. Successive uses of this argument allows the arrivals to node k to be expressed as a function of the arrivals to the system (i.e. to node 1).

$$\alpha_k(t) = \delta_{k-1}(t) \tag{2.7}$$

$$= \int_0^t \alpha_1(u) h_{k-1}(t-u) du. \tag{2.8}$$

where $h_{k-1}(t)$ is the probability a patient spends a combined time t at nodes 1 to $k-1$.

2.4.4 Demand at each node

Using the departure rate function Massey and Whitt are able to compute the mean demand of node k in a $(M_t/G/\infty)^N/D$ network and subsequently, by adding up the mean demand of each equivalent node across all the deterministic route networks, compute the mean demand of that node in the $(M_t/G/\infty)^{N_1}/M$ network.

2.5 Networks of infinite server queues: health-care applications

Utley et al. (2009) expands on the analytical methods developed for single node systems in Gallivan and Utley (2005) and Utley et al. (2003) to systems where patients move to a number of different states. Utley et al. (2009) describe a multi-state flow system as a rooted directed tree - a system where patients have a single point of entry and each other state is accessible by exactly one directed path. Patient arrivals are modelled using a time-varying general arrival distribution and general patient length of stay distribution. In this case they provide an example

of an application of their results to patients suffering from mental health problems such as anxiety or depression. Patients are categorised into one of seven states ranging from low/high intensity therapy to leaving the system either by dropping out or completing therapy. Using their analytical model Utley et al. estimate the mean and variance of occupancy at each of the states.

Helm and Van Oyen (2014) and Izady and Worthington (2011) provide two quite different applications of the theory on networks of infinite server queues in healthcare settings. Helm and Van Oyen (2014) seek to model an entire hospital with multiple wards, where elective and emergency patients can move between wards over time. They compute the mean demand of each ward and implement this information into optimisations aiming to determine optimal elective admission plans. Izady and Worthington (2011) use the mean demand (or offered load) of a service station k in an $(M_t/G/\infty)^{N_1}/M$ network to set hour by hour staffing levels while achieving a certain target delay probability.

Helm and Van Oyen (2014) provide an in depth investigation into modelling a hospital with multiple wards as a network of infinite server queues. They state that their work has contributed by developing analytical census modelling methods, as opposed to simulation-based methods, to characterise patient flows and schedule elective patient admissions. They also claim that their work is able to solve the scheduling portion of the Hospital Admission Scheduling and Control (HASC) problem, targeting the problem of census variability through the improved management of elective patient admissions in the form of optimisations. This paper is therefore compared in depth with Gallivan and Utley's schedule optimisation approach later in Section 2.6. In this section we provide an overview of both Helm and Van Oyen (2014)'s and Izady and Worthington (2011)'s healthcare applications of the theory on networks of infinite server queues.

2.5.1 Coping with infinite servers

Helm and Van Oyen (2014) and Izady and Worthington (2011) use infinite server models as these types of models provide closed form formulae for the mean demand at individual wards in the network (Equation (2.6)), similarly Utley et al. (2009) derive formulae for mean and variance of demand at individual nodes in their rooted directed tree system. These convenient results for the mean and variance are in part due to patients not interacting with one another as they do not share resources, and hence allow modellers to incorporate multiple patient types while computing the total mean demand of a ward.

The drawback of infinite server models is the often unrealistic assumption of sufficient capacity always being available at each ward i.e. no patient is ever turned away from the ward. Utley et al. (2009) use the infinite server model to compute the distribution of "demand" at each state in a system referring to the potential demand not subject to capacity constraints. They state this approach can be used to gain insights into the behaviour of a system that has capacity constraints.

Helm and Van Oyen (2014) discuss the impact of off-unit patients (patients placed on incorrect wards) and bed blockages (when patients are unable to enter the system due to the total hospital census exceeding the hospital capacity) and the negative results this has on patient lengths of stay and census variability. They address this issue by imposing capacity requirements at the optimisation stage of their methodology.

In Izady and Worthington (2011) the mean demand of the network infinite server model is necessary in the square root staffing law proposed by Jennings et al. (1996). Consider an $M_t/G/s(t)$ model where $s(t)$ represents number of servers at time t. Then to achieve some delay probability, α, Jennings et al. (1996) proposed the square root staffing function:

$$s_k(t) = \lceil m_k(t) + \beta\sqrt{m_k(t)} \rceil,$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x, β is a quality of service parameter and $m_k(t)$ is the time dependent offered load of node k. In this case $m_k(t)$ is estimated by the mean demand of node k in the associated $(M_t/G/\infty)^{N_1}/M$ model (Equation 2.6). The goal of estimating the required staffing levels accounts for capacity limits in the system.

2.5.2 Modelling hospitals with infinite server networks

Helm and Van Oyen (2014) and Izady and Worthington (2011) draw directly from the theory in Massey and Whitt (1993) to model their healthcare systems as networks of infinite server queues. Helm and Van Oyen model the bed demand of emergency and elective patient types using the Poisson Arrival Location Model (PALM) (Massey and Whitt, 1993) and Deterministic Arrival Location Model (DALM) models respectively. The DALM model represents an extension of the PALM model where arrivals are assumed to occur deterministically. With these respective models for each patient type Helm and Van Oyen are able to compute the contribution of each patient type to the mean demand of each ward, and consequently (due to the independence of patients in the infinite server model) the total mean demand of each ward. They also provide results for the variance of bed demand at both ward and total hospital level but do not apply this information at any stage.

When computing the mean demand of node k, $m_k(t)$, in the $(M_t/G/\infty)^{N_1}/M$ network, Izady and Worthington (2011) adopt the decomposition method proposed by Massey and Whitt (1993). They express the original Markov route network as a set of tandem node networks with deterministic routes, $(M_t/G/\infty)^N/D$. Then with the use of the departure rate function (Equation (2.8)) they are able to compute the mean demand of each node of their network. Unlike Helm and Van Oyen (2014), Izady and Worthington (2011) use this information to obtain the workload of certain types of staff, namely doctors and emergency nurse practitioners (ENPs).

Utley et al. (2009) take a different approach where the type of system they consider is a rooted directed tree graph. In this type of directed path model patients arrive at a system at one point of entry (the root) and access each other state by exactly one directed path. By defining the probability a patient is still in some state i, b time units after entering state i, known as state persistence probabilities and the probability a patient leaves state i and enters state j after being in state i for b time units, known as transfer probabilities, Utley et al. are able to derive formulae for the mean and variance of demand at some state i, T time units into the modelled period.

In some sense the concepts applied in Helm and Van Oyen (2014) and Izady and Worthington (2011) are the same, calculating the mean demand of a node to indicate the number of patients present at that node (assuming sufficient capacity is always available). But where Helm and Van Oyen (2014) take this information to indicate the beds required at the ward, Izady and Worthington (2011) are interested in the number of staff required. This reflects their differing objectives. Helm and Van Oyen (2014) aim to find "optimal" admission schedules to maximise the throughput of elective patients in one case whilst minimising the number of blockages and off-unit patients in the other. Izady and Worthington (2011) require the mean demand as a component of the square root staffing law and, with the use of simulation, aim to achieve staffing levels such that the patient discharge percentage is below some (government defined) level.

2.6 Comparison of methods for modelling patient flow

Gallivan and Utley (2005) and Helm and Van Oyen (2014) both seek to develop methods of booking scheduled patients in treatment centres (TCs), and in a hospital respectively. In order to improve elective patient scheduling, understanding bed demand is important as this often acts as a limiting factor in patient flows,

increasing blockages and waiting times.

It is worth noting that while both papers focus on booking elective admissions efficiently the models still need to account for emergency patients, which have a significant impact on bed demand. Both papers use infinite server models of patient flows to estimate the potential bed requirements in the system. Furthermore they both take the cases of elective and emergency patients separately, characterising the demand of each type of patient first before combining them.

A third related paper by Bekker and Koeleman (2011) shares the same objective of booking scheduled admissions in an efficient manner and aims to minimise the variation in occupancy levels by analysing the impact that the variability in the admissions and lengths of stay have on the required capacity. However this paper differs from Gallivan and Utley (2005) and Helm and Van Oyen (2014) in that the variance of bed demand is not computed.

We focus on these three papers (Gallivan and Utley (2005), Helm and Van Oyen (2014), and Bekker and Koeleman (2011)) as they use infinite server models to characterise patient flows and devise optimisation schemes to determine elective admission schedules. In this section we describe how the model outlined in Helm and Van Oyen (2014) is a generalisation of the model in Gallivan and Utley (2005). Furthermore we compare the approach used by Bekker and Koeleman with the other two papers.

2.6.1 Mean bed demand

Both papers assume the admission plans are cyclic; generally each cycle is a week long corresponding to a weekly schedule.

We begin with an outline of Gallivan and Utley's model, carefully defining the mean bed demand on day d of the cycle. Note that Gallivan and Utley assume admissions plans are cyclic where in general each cycle is a week long corresponding

to a weekly schedule.

$$\mu_d = \sum_{h=1}^{H} \sum_{i=1}^{C} N_{h,i} \sum_{w=0}^{\infty} p^h_{i,(wC+d-i)} + \sum_{i=1}^{C} \sum_{w=0}^{\infty} N_{0,i} p^0_{i,(wC+d-i)}, \tag{2.9}$$

where

- the h index indicates patient type, and H is the number of elective patient types,
- i indicates the day of arrival ($i = 1, 2, ..., C$), and C is the cycle length (usually a week, i.e. $C = 7$),
- w indicates the week or cycle,
- $N_{h,i}$ is the elective admission plan for patient type h on day i for $1 \leq h \leq H$,
- $p^h_{i,j}$ denotes the probability a patient of type h who is booked at time i is resident j days later (probability distributions are associated with each patient type),
- $N_{0,i}$ is the mean number of emergency arrivals on day i where $h = 0$ indicates the emergency patient type, and
- $p^0_{i,j}$ denotes the probability an emergency patient arrives at time i is still resident j days later.

In order to interpret the formula for mean bed demand we first consider a single elective patient type h. The contribution to mean bed demand from previous admissions of patients of type h is:

$$\sum_{i=1}^{C} N_{h,i} \sum_{w=0}^{\infty} p^h_{i,(wC+d-i)}$$

i.e. the booked admission on day of week i, $N_{h,i}$, multiplied by the probability the patient is still residing in the system on day d given they were admitted w weeks ago on day i of that week, $p^h_{i,(wC+d-i)}$. The same principle is applied to the

emergency patient type. The contribution from every patient type is calculated separately and summed together to determine the total mean bed demand.

Helm and Van Oyen take the same approach when computing the mean bed demand. They consider elective and emergency patient types separately. The number of elective admissions are assumed to be deterministic and emergency arrivals follow a non-homogeneous Poisson process.

The main difference between the models is the network nature of the hospital system which Helm and Van Oyen (2014) choose to model. By choosing to model a hospital with a network of wards, tracking patients and their resource requirements is an important building block for Helm and Van Oyen (2014). They achieve this by developing the Patient Temporal Resource Requirements (PATTERN) Stochastic Location Process model. This models the expected number of patients at each ward over time (which depends on each patient type's distribution of time spent in the system) and describes their flow through the network along with their resource requirements. Helm and Van Oyen extend their stochastic location model and incorporate Poisson and deterministic arrivals to summarise the arrival and resource requirements of emergency and elective patients respectively, known as the PATTERN PALM and DALM models respectively.

In the case of Gallivan and Utley (2005) only a simplified PATTERN PALM and DALM model is required. Since patients can only either be present in the system or not (i.e. discharged) the location model is restricted to a single ward case. The PATTERN PALM and DALM models, after being simplified to the single ward case, describe the expected number of patients in the system over time which depends on each patient type's length of stay persistence distributions which is the same as the Gallivan and Utley (2005) model.

2.6.2 Variance of bed demand

A benefit of both papers characterising their respective healthcare systems analytically is that they are able to derive formulae for the mean and variance of bed

demand. They share a similar approach by calculating the variance of elective and emergency patient bed demand separately.

Gallivan and Utley define the variance of bed demand on day d:

$$\sigma_d^2 = \sum_{h=1}^{H}\sum_{i=1}^{C} N_{h,i} \sum_{w=0}^{\infty}\sum_{w=0}^{\infty} p^h_{i,(wC+d-i)}(1 - p^h_{i,(wC+d-i)}) + \sum_{i=1}^{C}\sum_{w=0}^{\infty} N_{0,i} p^0_{i,(wC+d-i)}, \tag{2.10}$$

where the first element indicates the contribution from elective patients and the second the contribution from emergency patients.

Helm and Van Oyen's formulae for the variance of elective and emergency bed demand are extensions of Gallivan and Utley's. See Appendix A for explicit steps showing how the formulae for the elective and emergency patient demands in Helm and Van Oyen (2014) can be simplified to the Gallivan and Utley (2005) case.

Note that the contributions of emergency patients to both the mean and variance of bed demand (the second elements of the right-hand terms of Equations (2.9) and (2.10)) is the same in both cases. This stems from the fact that the number of beds required by emergency cases is Poisson distributed (Equation (2.4)), hence the mean and variance are the same.

While Gallivan and Utley are able to derive a formula for the variance of bed demand, they do not exploit this information when constructing a booking schedule at the optimisation stage of their method. Instead 90th and 95th percentiles for bed demand are calculated and used in analysing an example scenario. Similarly Helm and Van Oyen do not use the variance of the hospital census or ward census at the optimisation stage of the process.

2.6.3 Relationship with the Bekker and Koeleman (2011) model

Bekker and Koeleman (2011) present a third model which shares a number of similarities with Gallivan and Utley (2005) and Helm and Van Oyen (2014). The

aim of their model is to determine an optimal booked admission schedule to regulate the occupancy pattern. They use an infinite server model to help construct a quadratic programming model which is designed to minimise the squared error between a target load and the average load per day.

Bekker and Koeleman describe the state of the healthcare system by calculating the "offered load" which is the average number of patients present in the case of sufficient capacity. This is equivalent to the mean bed demand used in Gallivan and Utley (2005) which shares the same "sufficient capacity" assumption (by using infinite servers all patient bookings are honoured and emergency patients allowed to enter the system). Furthermore the Bekker and Koeleman (2011) model incorporates cyclic arrival patterns (i.e. arrivals changing each day of the week) for both emergency arrivals (characterised by the Poisson distribution) and elective arrivals (general arrival process), multiple patient types and different length of stay distributions for each type.

Determining the elective admission schedule is broken down into two stages. The first stage is to calculate some target load which involves consulting hospital managers. This stage accounts for limited capacity by using blocking probabilities to help inform the choice of target load. This step shares some similarities to Helm and Van Oyen's model which estimates the expected blockages and off-unit patient census and includes these elements in the optimisation stage of the model.

With a chosen target load, Bekker and Koeleman (2011) present a quadratic program to determine an optimal admission schedule. Bekker and Koeleman emphasise the importance of reducing variability in the healthcare system, and how stabilising the census level can improve the state of the healthcare system. With this in mind their objective function is to minimise the deviation of the offered load from the target load i.e. stabilising the census level by encouraging the bed census to be as close to some target load as possible. This objective reflects the difference between Bekker and Koeleman (2011) with the models in Gallivan and Utley (2005) and Helm and Van Oyen (2014). Helm and Van Oyen's optimisation

places more focus on maximising the elective admissions and reducing blockages as opposed to reducing variation in the hospital census. Furthermore Bekker and Koeleman explain they could have modelled admission schedule using a linear program (like in Gallivan and Utley (2005)) but choose to use a quadratic program. The reasoning behind this is their belief that the consequences of departing from the target load do not increase linearly as the size of the deviation increases. However, like the other two papers, Bekker and Koeleman do not use the variance of bed demand at any stage.

2.6.4 Summary and limitations of optimisation procedures

When it comes to the optimisation the papers, Gallivan and Utley (2005) and Helm and Van Oyen (2014), diverge slightly. Gallivan and Utley opt to minimise a dummy variable and interpret the results. In this case even after determining an optimal admission schedule the mean bed demand may still exceed the maximum capacity. On the other hand, Helm and Van Oyen incorporate bed-blockages and off-unit patients to reflect the limited capacity of the hospital wards before applying an optimisation.

These two different approaches at the end reflect the fact that Gallivan and Utley (2005) seek to illustrate the potential for formulating bed demand analytically in order to apply an optimisation. While on the other hand Helm and Van Oyen (2014) choose to show a specific method which they develop and felt had solved the scheduling problem optimally.

Both optimisations however, share the same weakness. There remains the issue of conflicting objectives within a hospital which impacts on the idea of an "optimal" elective patient schedule. For example Helm and Van Oyen propose two optimisations, corresponding to each objective: maximising the elective patient throughput and minimising the number of blockages. Each optimisation will not necessarily produce the same elective admission schedule nor will it necessarily be clear which solution is preferred; expert opinion and hospital managers

would influence the final admission schedule. The range of concerns and objectives for hospital decision makers means that there is likely to always be some level of subjectiveness when finding the best admission plan. Furthermore neither case addresses using the variance of bed demand in their optimisation procedures (though Gallivan and Utley demonstrate using the variance of bed demand in analysing their optimal admission plan).

The model presented in Bekker and Koeleman (2011) shares a number of similarities with Gallivan and Utley (2005) and Helm and Van Oyen (2014). In particular they all use infinite server queues when characterising the healthcare system in order to determine the potential number of occupied beds while incorporating a mixture of elective and emergency patient types. The optimisation in Bekker and Koeleman (2011) differs because of the shift in objective from maximising the elective admissions and minimising the number of blockages (Helm and Van Oyen, 2014) to reducing the variability in the census.

However Bekker and Koeleman (2011) also shares the same shortcomings as Gallivan and Utley (2005) and Helm and Van Oyen (2014). The concept of an "optimal" elective schedule remains subjective and lacks the implementation of the variance of bed demand.

2.7 Other uses of infinite server models

Having identified the lack of use of the variance of bed demand and the restrictions of optimisation, this raises the question of whether there are other methods for applying infinite server models to improve healthcare systems? One traditional approach is to use the infinite server models as "what if" models, to simply inform decision makers about the consequences of possible decisions, without trying to advise what is "optimal". An example of this type of analysis is in a report to Launceston General Hospital (LGH), which is attached in Appendix B

Another possibility is to use infinite server models to provide a "fair" basis on which to compare the performance of hospitals, i.e. to underpin performance

indicators. The following Section briefly introduces performance indicators, where we discuss how they inform modellers and decision makers on the performance of their systems. They offer an alternative approach to using the mean and variance of bed demand beyond optimisation. This idea forms the basis for constructing our own model-based performance indicators to measure hospital performance in Chapters 4 and 5.

2.8 Performance indicators

Performance indicators are statistics which provide insight into the behaviour and performance of some system. It is important to recognise that performance indicators do not explain the entire performance story, rather they indicate, they point to where performance seems to be good, poor, average, improving, static or declining (Pidd, 2012). They are useful because they summarise performance and allow comparison through time and between units.

In some sense performance indicators can be likened to models in that they are simplified representations of some real world system and it is not always possible, or even desirable, to incorporate every feature of a system into a performance indicator. Some simplification is necessary because the systems being summarised are complex. For example healthcare systems are multi-dimensional in terms of their objectives be it financial constraints which must be adhered to, a limited number of resources such as beds or staff, or maintaining patient quality of care. Like models, performance indicators are designed with a purpose in mind therefore its intended use is an important consideration when making simplifications and assumptions when constructing performance indicators.

The role of performance indicators in informing users and helping them understand the performance of systems is not without its drawbacks. It is important for users to understand what exactly the performance indicator is used for and what simplifications have been made otherwise there is a danger of misinterpreting them especially if a user fails to understand the organisation of the system of interest.

Furthermore performance measurement may encourage game playing, increase internal administration and block innovation. There is a danger that individuals focus on improving performance measures which can punish good performance. With this in mind we must always remind ourselves that performance indicators simply indicate and over-interpreting or placing too much emphasis on them can lead to the detriment of the performance of the system.

Performance indicators potentially serve a very useful purpose for healthcare systems. They are able to capture and measure the performance of complex systems, such as hospitals which often share a variety of conflicting objectives, and allow users to gain some understanding and compare hospital performance.

2.9 Discussion

In this chapter we establish the main results of single node infinite server models, addressing homogeneous and nonhomogeneous arrivals to infinite server queues, as well as networks of infinite server queues. Compared to other approaches used in healthcare modelling such as simulation, we motivate the use of infinite server models as they typically cost less, are simpler, easier to implement, and provide more generic results. The main aim of this thesis is therefore to investigate the potential to extend the use of infinite server queueing models in healthcare modelling.

In Chapter 3 we first look at the theoretical aspects of this issue. An important result from infinite server models is expressing the distribution of the number of customers in the system, in particular having tractable formulae for the mean and variance of demand. We investigate new theory by relaxing the assumptions placed on the arrivals but still expressing the mean and variance of bed demand in closed formulae. We also consider the complex methods used by Massey and Whitt (1993) when determining the mean demand of individual nodes in networks of infinite server queues, and offer an alternative and simpler approach.

In the remainder of the thesis we seek alternative uses of the mean and variance

of bed demand in terms of improving hospital decision making. While optimisation approaches offer one way of potentially improving hospital efficiency they are limited in their ability to capture the range of priorities faced by healthcare decision makers such as budgeting constraints, staffing hours, operating theatre availability, etc. Optimisations are restricted by our ability to accurately define these constraints and find a suitable solution for improving the hospital, in addition a secondary issue for more complex optimisations designed to incorporate more of these constraints have to consider computer time problems.

In both Gallivan and Utley (2005) and Helm and Van Oyen (2014) the mean and variance of bed demand are expressed explicitly, with the variance of bed demand identified as an important issue in good hospital management, but not really pursued. An important feature of our work is making use of the variance of bed demand. There are dangers of focusing solely on the mean bed demand, with healthcare decision makers often interested in the peaks and troughs of demand just as much as the overall level. We consider alternative approaches of using the formulae for the mean and variance of bed demand in the form of model-based performance indicators (Chapters 4, 5 and 6).

Chapter 3

Consolidating, simplifying and extending results for single node and networks of infinite server queues

In the previous chapter the theory and important results for infinite server models were summarised. This chapter focuses on two types of systems; single node infinite server queues and networks of infinite server queues. Our interest in these types of systems stems from a healthcare perspective, they can be used to model hospitals and other healthcare systems to help inform decision makers by predicting hospital performance.

Our interest in this chapter is how the mean and variance of bed demand can be calculated by modelling a healthcare system as a single node infinite server system or as a network of infinite server queues. Characterising a system using an infinite server model has two important modelling advantages. Firstly contributions to demand from different patient types are independent of one another since patient types do not compete for resources (e.g. beds) and so they do not interact or interfere with the flow of other patients in the system. As a consequence infinite

server systems allow us to express the mean and variance of bed demand analytically. Such formulae are potentially much more insightful and convenient than simulation models, and their application in optimisation models to construct elective admission schedules have already been described (Gallivan and Utley (2005), Helm and Van Oyen (2014)).

Gallivan and Utley (2005) develop a discrete time single node infinite server system in a healthcare setting to compute the mean and variance of bed demand followed by implementing an optimisation to determine an elective inpatient admission schedule. Massey and Whitt (1993) present a method of calculating the mean demand of individual queues in a continuous time network of infinite server queues with time dependent Poisson arrivals which forms the basis of the methods used in Helm and Van Oyen (2014). In this chapter we therefore first consolidate and where possible simplify the results for the mean and variance of bed demand from Gallivan and Utley (2005) and Helm and Van Oyen (2014). We also show that our results are equivalent to Massey and Whitt (1993)'s but our approach is much simpler and easier to understand.

Furthermore we extend the present theory, using ideas and concepts shown in Gallivan and Utley (2005) and Massey and Whitt (1993) in order to relax the assumptions placed on the arrival distributions and incorporate a general arrival distribution and multiple patient types. This differs from the Binomial or deterministic arrivals used in Gallivan and Utley (2005) and Helm and Van Oyen (2014) for elective patients respectively.

This chapter begins with a summary of the two key papers (Section 3.1), Gallivan and Utley (2005) and Massey and Whitt (1993), highlighting the main results and some important ideas. We then present new results for the mean and variance of bed demand in a single node infinite server system using a general arrival distribution (Section 3.2). An extension with corresponding formulae for the mean and variance of bed demand while accounting for multiple patient types is shown (Section 3.3). Interesting special cases of these results using Poisson arrivals, deter-

ministic arrivals and residual patients are presented, showing how our framework is able to incorporate these scenarios (Sections 3.4, 3.5 and 3.6). Then we introduce an easy-to-use approach for calculating the mean bed demand in networks of infinite server queues with a general arrival distribution and compare our result with that shown in Massey and Whitt (1993) (Section 3.7). Next we consider the distribution of the number of patients in the system when using a general arrival distribution (Section 3.8). Finally provide a summary of the chapter.

3.1 Existing results for single node and networks of infinite server queues

Before any new results are presented we first remind the reader of the existing theory relating to single node and networks of infinite server queues. In particular we highlight the key results and concepts from Gallivan and Utley (2005) and Massey and Whitt (1993), some of which we adopt in order to extend and broaden the current theory. Both papers address calculating the mean and variance of demand using infinite server queues, in the single node case (Gallivan and Utley, 2005) and networks of queues (Massey and Whitt, 1993).

Gallivan and Utley develop their methods in the context of booking elective patient admissions into a treatment centre by calculating the mean and variance of bed demand in the system where arrivals either occur following some Binomial distribution (elective patients) or Poisson distribution (emergency patients). The mean and variance of bed demand on day d are:

$$\mu_d = \sum_{h=1}^{H}\sum_{i=1}^{C} N_{h,i}\sum_{w=0}^{\infty} p^h_{i,(wC+d-i)} + \sum_{i=1}^{C}\sum_{w=0}^{\infty} N_{0,i}p^0_{i,(wC+d-i)},$$

$$\sigma^2_d = \sum_{h=1}^{H}\sum_{i=1}^{C} N_{h,i}\sum_{w=0}^{\infty}\sum_{w=0}^{\infty} p^h_{i,(wC+d-i)}(1-p^h_{i,(wC+d-i)}) + \sum_{i=1}^{C}\sum_{w=0}^{\infty} N_{0,i}p^0_{i,(wC+d-i)},$$

where

- the $1 \leq h \leq H$ index indicates elective patient type, and H is the number of elective patient types,
- i indicates the day of arrival, and C is the cycle length,
- w indicates the week or cycle,
- $N_{h,i}$ is the elective admission plan for patient type h on day i for $1 \leq h \leq H$,
- probability distributions are associated with each patient type; $p^h_{i,j}$ denotes the probability a patient of type h who is booked at time i is resident j days later,
- $N_{0,i}$ is the mean number of emergency arrivals on day i where $h = 0$ indicates the emergency patient type, and
- $p^0_{i,j}$ denotes the probability an emergency patient arrives at time i is still resident j days later.

Much of the apparent complexity of these formulae is due to their healthcare application which incorporates a number of important features which influence the bed demand. Elective and emergency patient types differ in the manner in which they arrive. Elective inpatient arrivals are deemed to follow a Binomial distribution and correspond to the first halves of both the formulae. Emergency patient arrivals are assumed to arrive randomly according to some Poisson distribution and correspond to the second halves of both the formulae. A cycle refers to an arrival pattern, usually a cycle length of 7 days is used referring to a weekly admission schedule. In both cases the arrivals are day dependent and the number of arrivals (or the rate of arrival in the emergency case) can differ depending on the day of the cycle.

Elective patients are further subdivided into patient types which are defined by their arrival patterns and length of stay distributions. Introducing multiple elective patient types is a relatively straightforward step due to the infinite server

assumption rendering the contribution to the demand from each patient type independent. Emergency patients however are not further subdivided into smaller patient types but are treated as a single patient type. Later we show how we are able to relax the assumptions placed on the arrival distributions of elective and emergency patients and allow the arrivals to take a general distribution while still being able to express the mean and variance of bed demand analytically. In addition we include multiple emergency patient types.

Gallivan and Utley's formulae work solely in discrete time. When measuring the demand placed on hospitals, the daily demand is generally the level of detail required of healthcare planners since hospitals work on a daily routine so working in discrete day time is a sensible choice. Also data may dictate working in discrete day intervals as arrival and length of stay information provided may only state which day a patient arrives or how many days they remain in the treatment centre.

Unlike Gallivan and Utley (2005), the formulae in Massey and Whitt (1993) are in continuous time. Because of our interest in healthcare settings, our interest is primarily in a discrete time version of the Massey and Whitt (1993) models. As will be seen later, the discrete time results and the continuous time results are closely related.

Massey and Whitt (1993) use infinite server models to calculate the mean demand of individual queues in a network of infinite servers rather than a single node case as in Gallivan and Utley (2005). The mean demand of queue i at time t is given as follows

$$m_i(t) = \int_{-\infty}^{t} \alpha_i(u) G_i^C(t-u) du$$

where

- $\alpha_i(u)$ is the arrival rate to node i at time u, and
- $G_i^C(.) = 1 - G_i(.)$ is the survivor function associated with the length of stay at node i.

Note that Massey and Whitt also show that the queue lengths at time t at each node, $\{Q_i(t) : i = 1, 2, .., N\}$, of a $(M_t/G/\infty)^N/M$ model are independent Poisson random variables. Therefore the queue length mean is equivalent to the queue length variance in this case.

Multiple patient types are considered each having their own arrival and length of stay distribution where all the arrivals are all assumed to follow a Poisson distribution. Markovian routing is also assumed. This means the sequence of queues for a patient type is determined by the initial queue in which they arrive and the transition probabilities, independent of the arrival and length of stay distributions (Massey and Whitt, 1993).

A key concept introduced in Massey and Whitt (1993), which we adopt is the idea of decomposing an $(M_t/G_t/\infty)^{N_1}/M$ model into a set of $(M_t/G_t/\infty)^N/D$ models where D indicates the deterministic routing of customers in the system. We begin by dividing customers in the original $(M_t/G_t/\infty)^{N_1}/M$ model by their route through the system. All customers which share the same deterministic route are grouped together, modelled as an $(M_t/G_t/\infty)^N/D$ network. And all customers are assigned an $(M_t/G_t/\infty)^N/D$ model according to their path through the system. As customers are independent of one another in infinite server models, the mean demand of a node in the $(M_t/G_t/\infty)^{N_1}/M$ network is the sum of mean demands of the same node from each of the $(M_t/G_t/\infty)^N/D$ models.

This decomposition idea simplifies the original problem of computing the mean demand of a node in the $(M_t/G_t/\infty)^{N_1}/M$ network to the point where we need only find the mean demand of a node in an $(M_t/G_t/\infty)^N/D$ model. Note that N may be greater than N_1 as customers may visit a node more than once.

In a healthcare context the original $(M_t/G_t/\infty)^{N_1}/M$ network can represent a healthcare system with patients of multiple types moving between wards where we are concerned with the mean bed demand. Each $(M_t/G_t/\infty)^N/D$ network represents a single patient type which has a fixed (deterministic) route through a set of tandem wards. Should a patient visit a ward more than once, each new visit

is treated as a separate ward.

A feature of Massey and Whitt's work on networks of infinite server queues is to assume the external arrival process is Poisson. After decomposing the network into a set of tandem node networks, this allows the arrival process at each (subsequent) node to be expressed analytically as a function of the departure rates of the preceding nodes (i.e. the queues the patient was present in prior to arriving at the queue of interest). Then using the departure rate function of the "previous" queue, we can find an expression for the arrival rate function of the queue we are interested in and given each queue has a length of stay distribution, Massey and Whitt are able (in theory) to compute the mean demand of any queue in a network of infinite servers queues. This concept of determining arrival rate functions using departure rate functions may seem intuitive but in practice it is not straightforward to find the expression for the arrival rate and implement it to compute the mean demand.

We therefore present a more intuitive approach for calculating the mean demand in networks of queues without the use of departure rate functions, which we also show to provide the same result as that in Massey and Whitt (1993).

The main idea in our intuitive approach is that the mean demand of any queue in a network is the difference between the total mean demand (the demand from all of the queues in the network) and the mean demand of all of the queues bar the queue of interest. This stems from the fact that a patient in the system must be present in one (and only one) of the queues.

3.2 Mean and variance of bed demand with a general arrival distribution

Gallivan and Utley (2005) present formulae for the mean and variance of bed demand of a single node system with multiple elective patient types and an emergency patient type. They assumed elective patients arrive according to some Bi-

nomial distribution (i.e. booked patients arrived according to some fixed schedule although there accounts for random non-arrivals) and emergency patient arrivals follow some Poisson distribution. This section relaxes the assumptions placed on the arrivals allowing a general distribution of daily arrivals, and shows how to compute the mean and variance of bed demand.

Theorem 3.2.1 *Let the number of patients in the system at discrete time t be X_t and let $X_{t,u}$ be the number of patients in the system at discrete time t given they arrived at time u for $u \leq t$. Let $A_u = j$ indicate that the number of arrivals at time u was j, for $j = 1, 2, ...$ and let $p_{j,u} = P(A_u = j)$. Then the mean and variance of bed demand at time t are:*

$$E(X_t) = \sum_{u=-\infty}^{t} E(A_u)s_{u,t}$$

$$var(X_t) = \sum_{u=-\infty}^{t} \left(E(A_u)s_{u,t}(1 - s_{u,t}) + E(A_u^2)s_{u,t}^2 - [E(A_u)s_{u,t}]^2\right)$$

where $s_{u,t}$ is the probability of a patient arriving at time u and surviving until at least time t.

Proof. A key concept of incorporating a general arrival distribution when computing the mean and variance of bed demand is the distribution of the number of patients in the system at time t, given they arrived at time u, $X_{t,u}$. Given a general arrival distribution and the length of stay distribution then the number of patients still in the system at time t (having arrived at time u) can be modelled as a mixture of Binomial distributions. Each Binomial in the mixture distribution represents some (fixed) number of arrivals with an associated probability of survival. Then the distribution of number of patients in the system at time t who

arrived at time u can be defined as

$$X_{t,u} = \begin{cases} \text{Binom}(1, s_{u,t}) & \text{if } A_u = 1 & p_{1,u} \\ \text{Binom}(2, s_{u,t}) & \text{if } A_u = 2 & p_{2,u} \\ \quad\vdots & \quad\vdots & \quad\vdots \\ \text{Binom}(K, s_{u,t}) & \text{if } A_u = K & p_{K,u} \end{cases}$$

where $s_{u,t}$ is the probability of a patient arriving at time u and surviving until at least time t.

Then X_t is made up of contributions from arrivals on past days, u, and have survived up to time t, $X_{t,u}$ for $u = t, t-1, ...$ which we can express as follows

$$X_t = X_{t,t} + X_{t,t-1} + ...$$

Then the mean number of patients in the system at time t is

$$\begin{aligned} E(X_t) &= E(X_{t,t} + X_{t,t-1} + ...) && (3.1)\\ &= E(X_{t,t}) + E(X_{t,t-1}) + ... && (3.2)\\ &= \sum_{u=-\infty}^{t} E(X_{t,u}) && (3.3)\\ &= \sum_{u=-\infty}^{t} \sum_{j=1}^{K} p_{j,u} E(X_{t,u} | A_u = j) && (3.4)\\ &= \sum_{u=-\infty}^{t} (p_{1,u} E(X_{t,u} | A_u = 1) + ... + p_{K,u} E(X_{t,u} | A_u = K)) && (3.5)\\ &= \sum_{u=-\infty}^{t} (p_{1,u} 1 s_{u,t} + p_{2,u} 2 s_{u,t} + ... + p_{K,u} K s_{u,t}) && (3.6)\\ &= \sum_{u=-\infty}^{t} (p_{1,u} 1 + p_{2,u} 2 + ... + p_{K,u} K) s_{u,t} && (3.7)\\ &= \sum_{u=-\infty}^{t} E(A_u) s_{u,t} \quad \text{as required,} && (3.8) \end{aligned}$$

where the step from Equation (3.3) to (3.4) follows from the simple definition:

$$E(X_{t,u}) = p_{1,u}E(X_{t,u}|A_u = 1) + ... + p_{K,u}E(X_{t,u}|A_u = K) \tag{3.9}$$

$$= \sum_{j=1}^{K} p_{j,u}E(X_{t,u}|A_u = j). \tag{3.10}$$

Similarly for the variance of bed demand:

$$\text{var}(X_t) = \text{var}(X_{t,t} + X_{t,t-1} + ...) \tag{3.11}$$

$$= \text{var}(X_{t,t}) + \text{var}(X_{t,t-1}) + ... \quad \text{by independence of patient arrivals between days} \tag{3.12}$$

$$= \sum_{u=-\infty}^{t} \text{var}(X_{t,u}). \tag{3.13}$$

Focusing on the individual terms in Equation (3.13) and using the law of total variance (Weiss et al., 2006) we write $\text{var}(X_{t,u})$ as

$$\text{var}(X_{t,u}) = E_{A_u}(\text{var}(X_{t,u}|A_u)) + \text{var}_{A_u}(E(X_{t,u}|A_u)). \tag{3.14}$$

Taking the first element of Equation (3.14) and using the law of total expectation (Weiss et al., 2006)

$$E_{A_u}(\text{var}(X_{t,u}|A_u)) = p_{1,u}\text{var}(X_{t,u}|A_u = 1) + ... + p_{K,u}\text{var}(X_{t,u}|A_u = K) \tag{3.15}$$

$$= \sum_{j=1}^{K} p_{j,u}\text{var}(X_{t,u}|A_u = j). \tag{3.16}$$

Now $X_{t,u}|A_u = j \sim \text{Binom}(j, s_{u,t})$. Therefore $\text{var}(X_{t,u}|A_u) = js_{u,t}(1 - s_{u,t})$.

And substituting into Equation (3.16) gives:

$$E_{A_u}(\text{var}(X_{t,u}|A_u)) = p_{1,u}\text{var}(X_{t,u}|A_u = 1) + ... + p_{K,u}\text{var}(X_{t,u}|A_u = K) \tag{3.17}$$

$$= p_{1,u}1s_{u,t}(1 - s_{u,t}) + ... + p_{K,u}Ks_{u,t}(1 - s_{u,t}) \tag{3.18}$$

$$= E(A_u)s_{u,t}(1 - s_{u,t}). \tag{3.19}$$

Taking the second term from Equation (3.14), using $\text{var}(Z) = E(Z^2) - E(Z)^2$:

$$\begin{aligned}\text{var}_{A_u}(E(X_{t,u}|A_u)) &= p_{1,u}E(X_{t,u}|A_u = 1)^2 + ... + p_{K,u}E(X_{t,u}|A_u = K)^2 \\ &\quad - [p_{1,u}E(X_{t,u}|A_u = 1) + ... + p_{K,u}E(X_{t,u}|A_u = K)]^2 \qquad (3.20) \\ &= p_{1,u}(1s_{u,t})^2 + ... + p_{K,u}(Ks_{u,t})^2 - [p_{1,u}1s_{u,t} + ... + p_{K,u}Ks_{u,t}]^2 \\ &\qquad (3.21) \\ &= E(A_u^2)s_{u,t}^2 - [E(A_u)s_{u,t}]^2 . \qquad (3.22)\end{aligned}$$

Finally substituting Equations (3.19) and (3.22) into Equation (3.13) the variance of bed demand at time t is

$$\text{var}(X_t) = \sum_{u=-\infty}^{t} \left(E(A_u)s_{u,t}(1 - s_{u,t}) + E(A_u^2)s_{u,t}^2 - [E(A_u)s_{u,t}]^2\right) \quad \text{as required.}$$

3.3 Extension: Multiple patient types

We have described how to calculate the mean and variance of bed demand of a single patient type. However this theory can be quite simply extended to the mean and variance of a system with multiple patient types. Under the infinite server assumption every patient type (and in fact every patient) is independent of every other patient type since they do not share resources (Gallivan and Utley, 2005). Therefore the total mean and variance of bed demand from H patient types is the sum of the means and variances of each patient type.

For $1 \leq h \leq H$, let $X_{t,u}^h$ be the number of patients of type h in the system at time t given they arrived at time u. Let $p_{j,u}^h$ be the probability j arrivals occur at time u from group h.

For $1 \leq h \leq H$, let A_u^h be the number of arrivals at time u of patients of type h. And let $s_{u,t}^h$ be the probability a patient of type h arrives at time u and survives

to at least time t. Then the mean and variance of bed demand at time t is

$$E(X_t) = \sum_{h=1}^{H} \sum_{u=-\infty}^{t} E(A_u^h) s_{u,t}^h$$
$$\text{var}(X_t) = \sum_{h=1}^{H} \sum_{u=-\infty}^{t} \left(E(A_u^h) s_{u,t}^h (1 - s_{u,t}^h) + E((A_u^h)^2)(s_{u,t}^h)^2 - \left[E(A_u^h) s_{u,t}^h \right]^2 \right).$$

3.4 Special case: Poisson arrivals

We may be interested in modelling arrivals using the Poisson distribution as is often the case with emergency arrivals which are considered to be unplanned and random. This result is equivalent to the emergency patient case presented in Gallivan and Utley (2005). In the infinite server system if the arrivals follow some Poisson distribution then the number in the system is also Poisson distributed (Brown and Ross (1969)).

Let $X_{t,u}^0$ be the number of emergency patients still in the system at time t given they arrived at time u. Then if the arrivals at time u are assumed to be Poisson distributed with mean and variance, A_u^0, then the mean and variance of the number of emergency patients in the system at time t, X_t^0, is:

$$E(X_t^0) = \sum_{u=-\infty}^{t} E(X_{t,u}^0) \tag{3.23}$$
$$= \sum_{u=-\infty}^{t} E(A_u^0) s_{u,t}^0, \tag{3.24}$$

where $s_{u,t}^0$ is the probability an emergency patient arrives at time u survives to at least time t.

3.5 Special case: Deterministic arrivals

For elective patients who arrive according to some fixed schedule we may choose to treat their arrivals as deterministic. Let $p_{j,u}$ be the probability j arrivals occur

at time u. Calculating the mean and variance of bed demand with deterministic arrivals can be achieved by setting all the $p_{j,u}$'s to 0, except for the one case relating to the known number of arrivals. So the number in the system at time t is:

$$X_{t,u} = \begin{cases} \text{Binom}(1, s_{u,t}) & \text{if } A_u = 1 & p_{1,u} = 0 \\ \text{Binom}(2, s_{u,t}) & \text{if } A_u = 2 & p_{2,u} = 0 \\ \vdots & \vdots & \vdots \\ \text{Binom}(D_u, s_{u,t}) & \text{if } A_u = D_u & p_{D_u,u} = 1 \\ \vdots & \vdots & \vdots \\ \text{Binom}(K, s_{u,t}) & \text{if } A_u = K & p_{K,u} = 0 \end{cases}$$

where D_u represents the deterministic number of arrivals on day u and this occurs with probability 1. This effectively means we model the number of patients in the system at time t ($X_{t,u}$) as a single Binomial distribution with D_u arrivals and its associated survival probability, $s_{u,t}$.

Under the deterministic arrivals scenario the mean bed demand is therefore

$$E(X_t) = \sum_{u=-\infty}^{t} \sum_{j=1}^{K} p_{j,u} E(X_{t,u}|A_u = j) \tag{3.25}$$

$$= \sum_{u=-\infty}^{t} p_{D_u,u} E(X_{t,u}|A_u = D_u) \tag{3.26}$$

$$= \sum_{u=-\infty}^{t} D_u s_{u,t}. \tag{3.27}$$

To calculate the contribution to the variance of bed demand at time t from arrivals at time u, we consider the first and second components of Equation (3.14) separately. The first element becomes

$$E_{A_u}(\text{var}(X_{t,u}|A_u)) = p_{D_u,u} \text{var}(X_{t,u}|A_u = D_u) \tag{3.28}$$

$$= D_u s_{u,t}(1 - s_{u,t}). \tag{3.29}$$

The second element of Equation (3.14) becomes

$$\text{var}_{A_u}(E(X_{t,u}|A_u)) = 0 \qquad \text{as } A_u \text{ does not vary.} \tag{3.30}$$

Thus the variance of total bed demand at time t is

$$\text{var}(X_t) = \sum_{u=-\infty}^{t} D_u s_{u,t}(1 - s_{u,t}). \tag{3.31}$$

which is equivalent to the result presented in Gallivan and Utley (2005).

3.6 Special case: Residual patients

Residual patients offer a simple extension of the existing infinite server theory but as far as we are aware they have not been addressed in other literature. In practice infinite server models may be used to forecast future patient numbers in circumstances where some patients already occupy beds (i.e. residual patients), elective patients are scheduled and emergency patients are unscheduled. By defining the residual patients as a separate patient type, with its own length of stay distribution we are able to calculate the mean and variance of bed demand of residual patients. The total mean and variance of bed demand is then the sum of the contributions from the elective and emergency patient types plus the contribution from the residual patients.

3.7 Networks of infinite server queues

Networks of infinite server queues present a more complex problem than the single node system. Networks are sets of queues which are connected i.e. patients can move from any queue to any other queue. Furthermore external arrivals can occur at any queue. We calculate the mean demand at individual nodes of networks of infinite server queues in a discrete time formulation of the problem with general

time-dependent arrival distributions. Massey and Whitt (1993) present a formula for calculating the mean demand at individual nodes of a network of infinite server queues with Poisson arrivals in a continuous time formulation. Furthermore using the Poisson arrival assumption Massey and Whitt (1993) show that the number of customers at each node are Poisson and independent. We present an alternative and much simpler method of calculating the mean demand for a general distribution of arrivals, which we show to be the same as the Massey and Whitt (1993) result for the case of Poisson arrivals.

In this section we compute the mean demand of any queue in a network when arrivals follow a general distribution. In particular express the mean bed demand at a node in an infinite server network as the difference of two single node models i.e. the same type of systems as presented in Section 3.2.

In practice we may be faced with network systems with patients moving through a variety of nodes and not all of them necessarily following the same route. We adopt the decomposition method presented in Massey and Whitt (1993) and described earlier in Section 3.1, where the network is broken down into a set of tandem node networks where each patient type follows some deterministic route. Hence the mean demand of a single node in a $(M_t/G_t/\infty)^{N_1}/M$ network is comprised of the sum of mean demands at the corresponding nodes in each of the $(M_t/G_t/\infty)^N/D$ networks.

We therefore provide an expression for the mean demand of a single patient type at a single node of a $(M_t/G/\infty)^N/D$ network. This network consists of N tandem nodes with arrivals occurring at only node 1 and move through all the subsequent nodes $1, 2, ..., k, ..., N$ before leaving the system at the final N^{th} node. Each node has its own associated length of stay distribution.

Theorem 3.7.1 *Consider a network of N tandem nodes with arrivals occurring at only node 1 before patients move through all the subsequent nodes $1, 2, ..., k, ..., N$ and leave the system at the final N^{th} node. Each node has its own associated length of stay distribution. Then the demand at node k at time t, X_t^k, has a mean which*

can be expressed as:

$$E(X_t^k) = \sum_{u=-\infty}^{t} \alpha_1(u)(H_k^C(t-u) - H_{k-1}^C(t-u)) \text{ for any } 2 \leq k \leq N. \quad (3.32)$$

For $k = 1$ the mean demand simplifies to:

$$E(X_t^1) = \sum_{u=-\infty}^{t} \alpha_1(u) H_1^C(t-u) \quad (3.33)$$

i.e. the mean demand of a single node case. Where

- $\alpha_1(u)$ *is the expected number of arrivals to the system i.e. arriving at node* 1*, at time* u*,*
- $H_j^C(w)$ *is the probability a patient's total length of stay at nodes* $1, 2, ..., j$ *is greater than* w *i.e. the patient is still in the system up to node* j *by time* w*.*

Proof. We first draw attention to an important distinction between the length of stay distribution of an individual node k, G_k (the cumulative distribution function of the length of stay distribution at node k), and the combined length of stay distribution of the first k nodes, H_k (i.e. the convolution of all the length of stay distributions up to node k). This notion of "patient stays in the system up to the k^{th} node" is important when understanding how the mean bed demand at node k is defined.

The proof is then based on two intuitive statements. The first is that in a deterministic tandem network:

X_t^k =number of patients entering the system (node 1) before time t who have not left node k by time t

− number of patients entering the system before time t who have not left node $k-1$ by time t

Hence

$$\begin{aligned} E(X_t^k) =& E(\text{number of patients entering system before time } t \text{ who have not left node } k \\ & \text{by time } t) \\ & - E(\text{number of patients entering system before time } t \text{ who have not left node } k-1 \\ & \text{by time } t). \end{aligned} \tag{3.34}$$

The second is to recognise that in a deterministic tandem network, combined service at the first k nodes (or first $k-1$ nodes) can be viewed as one single service in a single node infinite server system. Therefore by viewing service at nodes 1 to k (and nodes 1 to $k-1$) as a single service at one node we can use Theorem 3.2.1 to give us

$$\begin{aligned} & E(\text{number patients entering system before time } t \text{ who have not left node } k \text{ by time } t) \\ & = \sum_{u=-\infty}^{t} \alpha_1(u) H_k^C(t-u) \end{aligned} \tag{3.35}$$

and

$$\begin{aligned} & E(\text{number patients entering system before time } t \text{ who have not left node } k-1 \text{ by time } t) \\ & = \sum_{u=-\infty}^{t} \alpha_1(u) H_{k-1}^C(t-u) \end{aligned} \tag{3.36}$$

Hence substituting Equations (3.35) and (3.36) into Equation (3.34):

$$E(X_t^k) = \sum_{u=-\infty}^{t} \alpha_1(u)(H_k^C(t-u) - H_{k-1}^C(t-u)) \quad \text{as required.} \tag{3.37}$$

3.7.1 Special case: Poisson arrivals in a $(M_t/G/\infty)^{N_1}/M$ network

We now show how our formula for the mean bed demand (Equation (3.32)) is equivalent to that presented in Massey and Whitt (1993). To do so we first rewrite Equation (3.32), using two substitutions.

The second substitution uses the fact that a patient's combined service time, S, in nodes $1, 2, ..., k$ can be viewed as their combined service time at nodes $1, 2, ..., k-1$ plus their service time at node k, with probability distribution $g_k(.)$. Hence:

$$P(S \leq t-u) = H_k(t-u) \tag{3.38}$$

$$= \sum_{s=0}^{t-u} g_k(s) H_{k-1}(t-u-s) \tag{3.39}$$

$$= \sum_{v=u}^{t} g_k(t-v) H_{k-1}(v-u) \tag{3.40}$$

where Equation (3.40) is obtained using the substitution $s = t - v$.

The first definition is

$$H_k^C(t-u) - H_{k-1}^C(t-u) = H_{k-1}(t-u) - H_k(t-u)$$

using $H_k^C(w) = 1 - H_k(w)$.

Thus we can rewrite the mean bed demand at node k at time t, Equation (3.37):

$$E(X_t^k) = \sum_{u=-\infty}^{t} \alpha_1(u)[H_k^C(t-u) - H_{k-1}^C(t-u)] \tag{3.41}$$

$$= \sum_{u=-\infty}^{t} \alpha_1(u)[H_{k-1}(t-u) - H_k(t-u)] \tag{3.42}$$

$$= \sum_{u=-\infty}^{t} \alpha_1(u) \left[H_{k-1}(t-u) - \sum_{v=u}^{t} g_k(v-u) H_{k-1}(t-v) \right] \tag{3.43}$$

Before showing this result is the same as the equivalent result presented in

Massey and Whitt (1993) we first need to highlight the use of *discrete versus continuous time*. In our results we have focused on discrete time as our work has a healthcare application in mind. Massey and Whitt (1993) on the other hand present results in continuous time and therefore use integrals to represent the mean demand. However the same concepts for computing the mean bed demand in an infinite server queue in discrete time still apply in continuous time. The continuous time case can be seen as the natural limit of the discrete time case if we consider very fine time intervals. Therefore we can change the expression in Equation (3.43) for the mean bed demand from discrete notation to continuous time notation to match that of Massey and Whitt (1993):

$$E(X_t^k) = \int_{u=-\infty}^{t} \alpha_1(u) \left[H_{k-1}(t-u) - \int_{v=u}^{t} g_k(t-v) H_{k-1}(v-u) dv \right] du, \quad (3.44)$$

where $\alpha(u)$ is the instantaneous arrival rate function to the system i.e. arriving to node 1 at time u whereas in discrete time it is the expected number of arrivals per day.

In order to show the equivalence between our results and that shown in Massey and Whitt (1993) consider an $(M_t/G_t/\infty)^N/D$ network. The mean demand at node k of an $(M_t/G/\infty)^N/D$ network according to Massey and Whitt (1993) is

$$m_k(t) = \int_{v=-\infty}^{t} \alpha_k(v) G_k^C(t-v) dv \quad (3.45)$$

where

- $G_k^C(w) = 1 - G_k(w)$ is the probability the patient has not left node k by time w i.e. is the survivor function of node k at time w, and
- $\alpha_k(v)$ is the arrival rate to node k at time v, i.e. the departure rate from node $k-1$, $\delta_{k-1}(v)$.

Massey and Whitt (1993) show the queue lengths at time t, $Q_i(t)$, of an $(M_t/G_t/\infty)^{N_1}/M$ network are Poisson. The Poisson nature of the queue lengths

is due to the Poisson arrivals, allows the departure rate function of the $k-1$ node at time t, $\delta_{k-1}(t)$, to be expressed as a function of the arrival rate at node 1 as follows

$$\begin{aligned}
\alpha_k(t) &= \delta_{k-1}(t) && (3.46)\\
&= \int_0^\infty \alpha_1(t-u)dH_{k-1}(u) && (3.47)\\
&= \int_0^\infty \alpha_1(t-u)\frac{d}{du}H_{k-1}(u)du && (3.48)\\
&= \int_0^\infty \alpha_1(t-u)h_{k-1}(u)du && (3.49)\\
&= \int_{-\infty}^t \alpha_1(u)h_{k-1}(t-u)du. && (3.50)
\end{aligned}$$

where $h_{k-1}(w)$ is the probability a patient spends a combined time w at nodes 1 to $k-1$.

Hence substituting the departure rate function (Equation (3.50)) into the expression for the mean bed demand (Equation (3.45)) then the mean bed demand at node k is:

$$\begin{aligned}
m_k(t) &= \int_{v=-\infty}^t \alpha_k(v)G_k^C(t-v)dv && (3.51)\\
&= \int_{v=-\infty}^t \delta_{k-1}(v)G_k^C(t-v)dv && (3.52)\\
&= \int_{v=-\infty}^t \int_{u=-\infty}^v \alpha_1(u)h_{k-1}(v-u)G_k^C(t-v)dudv && (3.53)\\
&= \int_{u=-\infty}^t \alpha_1(u)\int_{v=u}^t h_{k-1}(v-u)G_k^C(t-v)dvdu \quad \text{(change order of integration)} && (3.54)\\
&= \int_{u=-\infty}^t \alpha_1(u)\left[H_{k-1}(t-u) - \int_{v=u}^t g_k(t-v)H_{k-1}(v-u)dv\right]du \quad \text{as required} && (3.55)
\end{aligned}$$

where the final step from Equation (3.54) to (3.55) holds using integration by

parts, i.e. :

$$\int_{v=u}^{t} h_{k-1}(v-u)G_k^C(t-v)dv \tag{3.56}$$

$$= \left[G_k^C(t-v)H_{k-1}(v-u)\right]_{v=u}^{t} - \int_{v=u}^{t} g_k(t-v)H_{k-1}(v-u)dv \tag{3.57}$$

$$= \left[G_k^C(0)H_{k-1}(t-u) - G_k^C(t-u)H_{k-1}(0)\right] - \int_{v=u}^{t} g_k(t-v)H_{k-1}(v-u)dv \tag{3.58}$$

$$= H_{k-1}(t-u) - \int_{v=u}^{t} g_k(t-v)H_{k-1}(v-u)dv. \tag{3.59}$$

3.8 Distribution of the number of patients

Whilst the mean and variance provides valuable information about bed demand, it would be useful if we could derive the full distribution of bed demand as well. Given Poisson arrivals we know the distribution of the number of patients at individual nodes of network systems and at single node infinite server systems are Poisson (Massey and Whitt (1993), Gallivan and Utley (2005)). However if arrivals are not assumed to be Poisson but instead are modelled by a general distribution then the distribution of the number of patients in the system is less straightforward.

In Section 3.2 we showed that the distribution of the number of patients arriving in the past at time u who survived to time t, $X_{t,u}$ is a mixture of Binomial distributions. Furthermore the number of patients in the system at time t, X_t, is made up of the sum of $X_{t,u}$ terms for $u = t, t-1, ...$, therefore the distribution of X_t is the sum of mixtures of Binomial distributions.

We now show under certain conditions that the sum of two or more mixtures of Binomial distributions is again a mixture of Binomial distributions. Consider a simple example summing two mixture of Binomial distributions with common survival probability s:

$$X = \begin{cases} \text{Binom}(a, s) & p_a \\ \text{Binom}(b, s) & p_b \end{cases}$$

$$Y = \begin{cases} \text{Binom}(c, s) & p_c \\ \text{Binom}(d, s) & p_d, \end{cases}$$

where $f_X(x) = p_a\binom{a}{x}s^x(1-s)^{a-x} + p_b\binom{b}{x}s^x(1-s)^{b-x}$ and $f_Y(y)$ is defined similarly.

Then if we are interested in $Z = X + Y$ we use a convolution to find the distribution for Z, i.e. :

$$\begin{aligned}
P(Z = z) &= P(X + Y = z) \\
&= f_Z(z) \\
&= \sum_{x=0}^{z} f_X(x) f_Y(z - x) \qquad \text{discrete convolution formula} \\
&= \sum_{x=0}^{z} \left[p_a \binom{a}{x} s^x (1-s)^{a-x} + p_b \binom{b}{x} s^x (1-s)^{b-x} \right] \cdot \\
&\left[p_c \binom{c}{z-x} s^{z-x} (1-s)^{c-(z-x)} + p_d \binom{d}{z-x} s^{z-x} (1-s)^{d-(z-x)} \right] \\
&= p_a p_c s^z (1-s)^{a+c-z} \sum_{x=0}^{z} \binom{a}{x}\binom{c}{z-x} + p_a p_d s^z (1-s)^{a+d-z} \sum_{x=0}^{z} \binom{a}{x}\binom{d}{z-x} \\
&+ p_b p_c s^z (1-s)^{b+c-z} \sum_{x=0}^{z} \binom{b}{x}\binom{c}{z-x} + p_b p_d s^z (1-s)^{b+d-z} \sum_{x=0}^{z} \binom{b}{x}\binom{d}{z-x} \\
&= p_a p_c s^z (1-s)^{a+c-z} \binom{a+c}{z} + p_a p_d s^z (1-s)^{a+d-z} \binom{a+d}{z} \\
&+ p_b p_c s^z (1-s)^{b+c-z} \binom{b+c}{z} + p_b p_d s^z (1-s)^{b+d-z} \binom{b+d}{z}.
\end{aligned}$$

This can be seen to be the density of a mixture of Binomial distributions:

$$Z = \begin{cases} \text{Binom}(a + c, s) & p_a p_c \\ \text{Binom}(a + d, s) & p_a p_d \\ \text{Binom}(b + c, s) & p_b p_c \\ \text{Binom}(b + d, s) & p_b p_d. \end{cases}$$

This argument can be generalised to summing any number of mixture of Binomial distributions each of which can consist of any number of Binomial distributions. Increasing the number of mixture of Binomial distributions or the number

of component Binomial distributions will simply increase the number of Binomial distributions that make up the resultant mixture distribution.

However in this analysis we have assumed the same survival probability s for all the Binomials. In our context this would imply the probability a patient remains in the system until day t is independent of how long ago they arrived, which is clearly a very unlikely scenario.

If the survival probabilities are not constant (which we would expect in general) then we are not able to express the sum of a mixture of Binomial distributions as another mixture of Binomial distributions. So in general we are not able to express the distribution of the number of patients in the system in a simple, analytical fashion.

3.9 Summary

This chapter consolidates existing results for the mean and variance of bed demand of single node infinite server queues and networks of infinite server queues from two key papers (Gallivan and Utley (2005), Massey and Whitt (1993)). We use key ideas from both papers, specifically that infinite servers lead to independent patient types in the system as they do not compete for resources, and the decomposition of a $(M_t/G/\infty)^{N_1}/M$ network into a set of tandem node networks with deterministic routes.

By adopting methods from these papers we simplify and extend the existing theory. Both a papers place assumptions on the arrival distributions of the patients; Poisson arrivals (both papers) and Binomial arrivals (Gallivan and Utley, 2005). In the single node case an important step in extending the theory to incorporate a general arrival distribution is showing that the number of patients surviving from day u until at least day t distributed as a mixture of Binomial distributions. As a result the mean and variance of bed demand on day t is simply the sum of the means and variances of mixtures of Binomial distributions. We show that deterministic and Poisson arrivals, and residual patients are special cases of

the general arrival distribution case.

In the network case Massey and Whitt (1993) express the departure rate function analytically and with this are able to compute the mean bed demand of each queue within the network. However deriving the departure rate function is not straightforward. We present a simpler, intuitive method for computing the mean bed demand by modelling it as the difference between the mean demand of two single node infinite server queues. Furthermore while Massey and Whitt (1993) assume Poisson arrivals our expression for mean demand allows any distribution of numbers of arrivals.

We also include multiple emergency patient types for the single node case (Gallivan and Utley (2005) only include multiple elective patient types and a single emergency type) and multiple patient types in general for the network case.

We also investigate the distribution of the number of patients in the system while assuming a general arrival distribution. However whilst we show that under certain conditions the number of patients is distributed as a mixture of Binomial distributions, the necessary assumption is not realistic for the bed demand problem.

Chapter 4

Model-based performance indicators for emergency work

4.1 Introduction

In Section 1.4.2 we introduced performance indicators as an alternative method of using infinite server models to help improve healthcare systems. The aim of this chapter and the next is to motivate, develop and demonstrate model-based performance indicators in the context of emergency inpatients, elective inpatients and total inpatients. We shall develop the model-based performance indicators in the context of 30 hospitals.

This chapter concentrates on the emergency work of 30 hospitals of varying sizes. Our interest in the emergency work of hospitals from around England is in the potential to use our analytical model as a means to compute model-based performance indicators. We make the distinction between traditional performance indicators and model-based performance indicators before demonstrating their use with a dataset from 30 hospitals. The Chapter 5 will then further develop and apply the same methods to elective inpatients and total inpatients.

With the performance indicators decision makers are better informed of the performance of the hospitals, highlighting cases where hospitals are performing

below the level we would expect and cases where they are performing well. By including hospitals of various sizes we investigate how hospital size impacts on the relative variability of emergency work.

Results for the emergency admissions and occupancies of the 30 hospitals are included in this chapter however the focus of our analysis and discussion lies with with emergency bed occupancy. The reason for drawing our attention to the bed demand is because hospital managers are able to influence bed occupancy unlike emergency admissions, for example managers can modify the bed occupancy by changing staffing levels and frequency of discharges. That is not to say the results based on the arrivals provides no insight, for instance emergency arrival performance measures could diagnose hospitals suffering from unusually high (or low) levels of variability. We would expect this variability to carry over into the occupancy, therefore any interpretation of the occupancy results should acknowledge the preceding arrival variability. However there are limitations when it comes to evaluating hospital performance based on arrivals with the management of hospitals in mind. Therefore it is sensible to evaluate hospital performance based on the emergency bed occupancy since it is, at least in part, affected by hospital management. We bare in mind the impact of emergency admissions but recognise that bed occupancy provides the most useful and interesting insights into hospital performance relating to management control.

When evaluating the performance of the 30 hospitals we focus our attention on the variability of bed demand as opposed to the mean bed demand. The difficulty of the average bed occupancy is that it is largely beyond the control of hospitals managers. If we subscribe to the idea that emergency admissions are outside of our control since they cannot be turned away or placed on a waiting list and patient lengths of stay are fixed (in that discharging patients early is generally unfeasible) then the average occupancy is determined. Therefore since the average emergency occupancy of any given hospital is outside the control of hospital managers when evaluating the performance of hospitals we do not centre our attention on the mean

emergency bed occupancy.

The variability of bed demand is a different matter to the mean bed demand when it concerns hospital management. Hospitals are tasked with meeting the day to day variations in demand and understanding the variability is crucial in good hospital management. Peaks in bed demand mean that managers should plan ahead by ensuring there is a sufficient reserve of bed stock available to avoid bed blockages. This is not a simple task when we consider managers must also be aware of their limited resources and the potential cost of over-stocking the reserve bed capacity. This careful balance highlights the importance of bed management and why the variability of emergency bed demand can be used to provide useful performance measures to evaluate hospitals.

We are interested in the variability of emergency admissions and bed occupancy to evaluate and compare the performance of hospitals of varying sizes. It is important to acknowledge that much variation is outside the control of hospitals. For example the nature of emergency admissions, which occur randomly will always cause some degree of variability which a hospital has little control over but must accommodate. On the other hand there are sources of variability within the control over hospitals such as the time it takes to receive diagnostic results (impacting on patient lengths of stay). The performance of any hospital dealing with emergency work will always be affected by sources of variation outside and within a hospital's control which must be considered when analysing hospital performance.

The analytical model used in this chapter is the same as the model presented in the previous chapters of the thesis. We shall summarise the key aspects of the analytical model and implement it on each of the 30 hospitals to compute the mean and variance of emergency bed demand of each hospital.

For each of the 30 hospitals we provide so called model-based performance indicators to evaluate how a hospital is performing against their expected performance according to our model. This means using our analytical model, which incorporates known sources of variability at each hospital such as random arrivals,

day-of-week effect of arrivals and patient length of stay distributions, to produce performance indicators. We compare each hospital's true performance, reflected by their observed results based on the data, against the performance indicators. We discuss how to interpret the performance indicators, managers may be tempted to view them as a strict threshold in which to identify under-performing hospitals based on emergency work, but we propose a more considered approach using the performance indicators to highlight which hospitals warrant further investigation.

The aim of this chapter is to explore how our analytical model can be used to produce model-based performance indicators for the emergency work of 30 hospitals. We investigate the scope for using the model-based performance indicators for evaluating and comparing the performance of these hospitals. In particular to measure hospital performance by the variability of their bed occupancy while accounting for known sources of variability in the form of patient arrivals and lengths of stay.

This chapter begins with a summary of the arrival and occupancy data and an overview of the analytical model used to produce the performance indicators. This is followed by a description of the chosen performance measures relating to the variability in admissions and occupancy, justifying why we have chosen them. The results section begins with a small example analysing hospital performing using the observed results only, demonstrating the limitations of this type of analysis without the use of modelled results. Then both observed and modelled results on the emergency admissions and bed occupancy are presented, including an extension of the earlier example but now incorporating the modelled results to illustrate how the performance indicators provide a more accurate description of hospital performance. We conclude with a discussion on the use of the analytical model to produce performance indicators for emergency work, highlighting what the performance indicators show and how this more informed description of hospitals can be used in understanding their performance.

4.2 Modelling Emergency Patients

4.2.1 The data

Data from 30 hospitals of varying sizes were obtained, this included daily emergency patient arrivals and daily numbers of beds occupied by emergency patients over the course of one year. More detailed data relating to patient arrival and departure times may be recorded by hospitals, though these data are typically not reported and hence not readily available to the modeller.

The 30 hospitals analysed in this chapter have been selected as a representative sample in terms of hospital size from a larger set of hospitals in England. The number of emergency admissions range from 22 admissions per day up to 175 admissions per day. The average number of emergency beds occupied range from 54 occupied beds per day up to 858 beds per day (Table 4.1). We exclude day cases from the data, focusing on patients who stayed at least one night and therefore required an inpatient bed.

Hospital number	Average arrivals per day	Average beds occupied per day	Average length of stay	Hospital number	Average arrivals per day	Average beds occupied per day	Average length of stay
1	175	703	4.0	16	72	353	4.9
2	105	688	6.5	17	68	496	7.3
3	125	566	4.5	18	60	261	4.4
4	125	542	4.3	19	56	184	3.3
5	103	586	5.7	20	49	231	4.7
6	116	504	4.4	21	48	272	5.7
7	106	858	8.1	22	48	272	5.7
8	86	517	6.0	23	38	187	4.9
9	101	545	5.4	24	31	54	1.8
10	81	394	4.9	25	22	96	4.4
11	96	608	6.3	26	104	720	6.9
12	85	441	5.2	27	86	481	5.6
13	58	354	6.1	28	86	494	5.8
14	85	352	4.2	29	65	322	5.0
15	72	338	4.7	30	44	189	4.3

Table 4.1: Summary of emergency arrivals, occupancy and length of stay at 30 hospitals.

4.2.2 The model

We use the same analytical model introduced in Section 3.1 of the thesis to compute the mean and variance of emergency bed demand. While we do not describe the model in detail here, we remind the reader of the main features of the model. Since we assume emergency arrivals occur randomly we characterise them using the Poisson distribution which captures the unexpected nature of emergency arrivals. The model also incorporates the day-of-week effect in emergency arrivals. We expect the number of emergency arrivals to vary depending upon the day of week, in particular the average number of arrivals per day during weekdays is greater than the average number of arrivals at weekends in all 30 hospitals. Emergency arrival rates for each day of the week are estimated using the data (for each hospital) and input this into our model.

The patients' lengths of stay are modelled by a geometric distribution for all 30 hospitals, as only the mean length of stay was available (see Table 4.1 for mean length of stay values). Other distributions can be used if the data is available, as the model only requires "survival probabilities". Also the geometric being a discrete distribution reflects our interest in the (whole) number of days a patient remains in the system.

Another important feature of the analytical model is the infinite server assumption. With infinite servers we assume that emergency admissions are never turned away which is typically the case, emergency departments almost always accepts all arrivals. Furthermore the infinite server assumption means the length of stay is independent of the number of arrivals and the number of beds occupied since patients cannot block one another and cause increased patient lengths of stay.

It should be noted that possible seasonality in arrivals has not been incorporated into the model. This could have been included, had we chosen to, by using different patterns of arrival rates each month in a similar fashion to the day-of-week effect. We refrained from including seasonality as this would add an unnecessary layer of complexity to our investigation.

4.3 Performance indicators

4.3.1 Traditional performance indicators

In order to compare and evaluate the performance of each hospital relative to its size, we have selected three measures of variability applied to both admissions and occupancy. For admissions the three performance measures used are:

- the coefficient of variation (the standard deviation of arrivals/mean arrivals (CoV)),
- the proportion of "busy days" to the total number of days where a "busy day" has been defined as a day when the number of arrivals was at least 25% greater than the mean number of arrivals, and
- the ratio of peak arrivals to the mean number of arrivals where the peak arrivals has been defined as the 95^{th} percentile of arrivals i.e. 95% of days was at or below this number of arrivals.

Since emergency admissions are beyond the control of hospital managers the performance measures relating to the admissions indicate the level of natural (and unavoidable) variation present at each hospital.

For bed occupancy we take a similar approach and use the equivalent three measures of variability as we did for the arrivals:

- the beds occupied coefficient of variation,
- the proportion of "busy days" where a "busy day" has been defined as a day when the number of beds occupied by emergency patients was at least 10% greater than the mean number of beds occupied, and
- the ratio of peak occupancy to the mean number of beds occupied where the peak occupancy is defined as the 95^{th} percentile of the beds occupied i.e. 95% of days was at or below this number of beds occupied.

4.3.2 Model-based performance indicators

For all 30 hospitals the mean and variance of the arrivals and occupancy is calculated using the available data (observed) and our analytical model (modelled). To compute the modelled occupancy results we use the model to calculate the mean and variance of bed demand. With the modelled mean and variance calculating the coefficient of variation was straightforward (standard deviation/mean). For the busy days and peak days we assume the number of emergency beds occupied is Poisson distributed with parameter equal to the mean bed demand. Given the distribution of emergency beds occupied we are able to find the the proportion of busy days to the total and the ratio of peak days to the mean.

In doing so we produce both observed and modelled results for each of the performance measures. In some sense the observed results provide a true reflection of the hospitals performance while the modelled results provide us with an indication of how we would expect each hospital to perform given the data.

4.4 Results

4.4.1 Evaluating hospital performance using traditional performance indicators

To demonstrate the importance of the analytical model in evaluating hospital performance we first consider the type of analysis available without any modelled results. Since the model accounts for the expected variability at each hospital the analysis without the modelled results will neglect to include this information, consequently we anticipate hospitals with large expected variability to be perceived as performing poorly and the opposite for hospitals with small expected variability.

We have chosen to focus on the beds occupied CoV to demonstrate the differences between using and not using the modelled results, though an equivalent analysis could have been conducted with the percentage of busy days or the peak

occupancy with similar results.

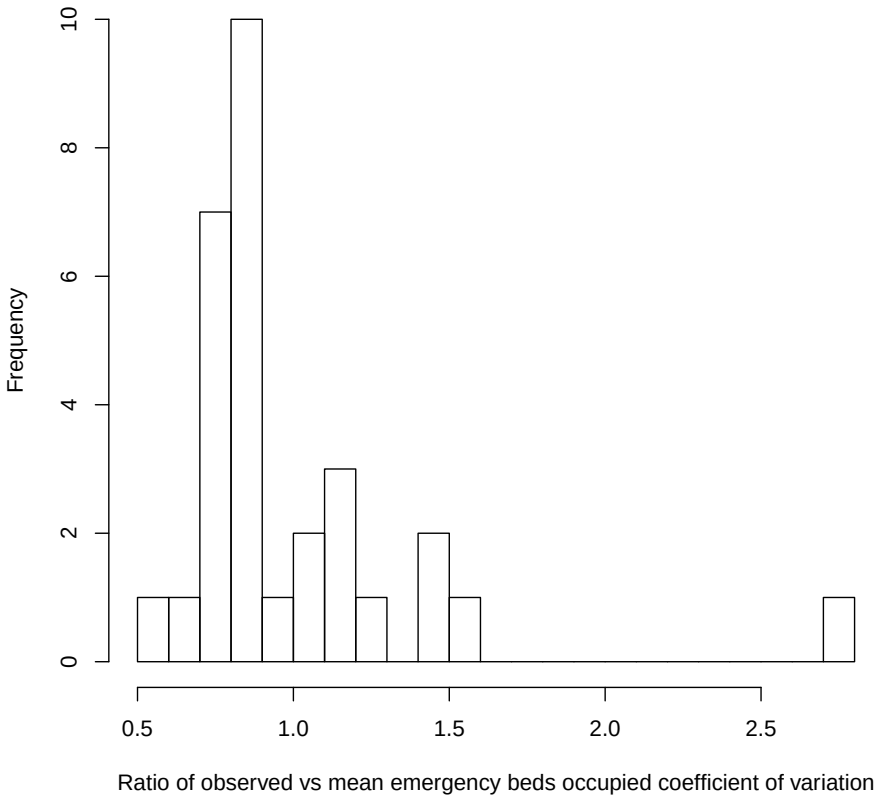


Figure 4.1: Histogram of the ratio of the observed over the mean beds occupied coefficient of variation.

The mean beds occupied CoV refers to the average CoV from all 30 hospitals and this is used to normalise the CoV results for comparing the hospitals. Each hospital has its own observed beds occupied CoV calculated from the data. So for each hospital we compute the ratio of the observed over the mean CoV.

The ratio of the observed over the mean CoV offers a simple approach for measuring the performance of each hospital using only the observed data (no modelled results are required). Hospitals with the largest ratios have the greatest variability, in terms of the beds occupied CoV, while those with the smallest ratios are be exhibiting the smallest beds occupied CoV. In either scenario, a large or small ratio would act as an indicator for further investigation to understand if the large/small CoV is due how the hospitals are being managed.

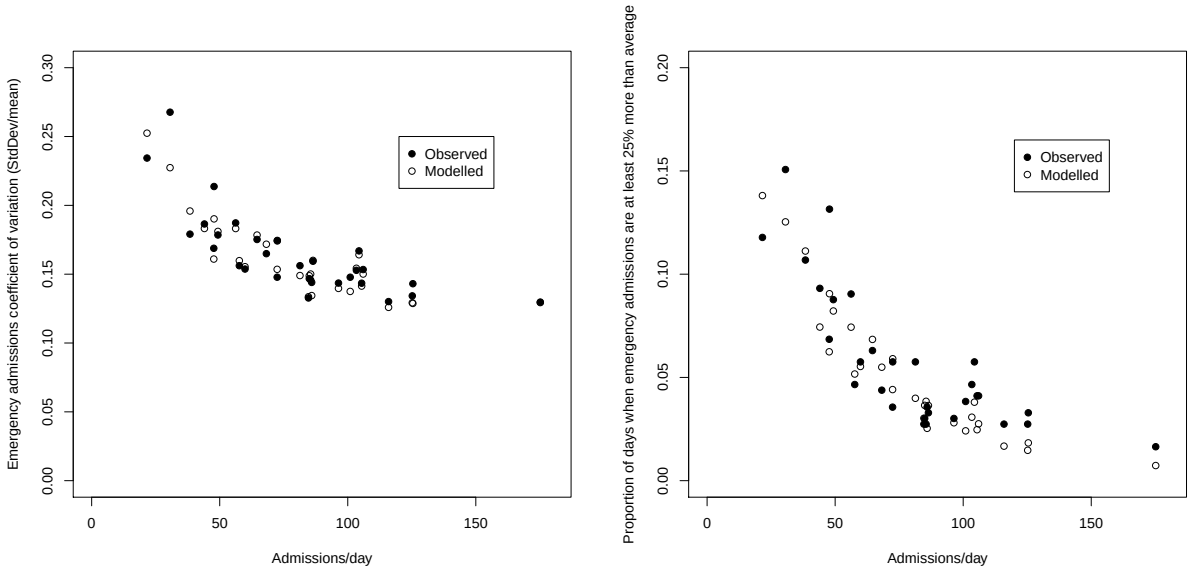
A histogram of the ratio of observed over the mean CoV, Figure 4.1, shows one hospital standing out, an outlier with a result of over 2.7. This is significantly larger than the remainder of the hospitals indicating this case is suffering from a much larger CoV in bed occupancy. Based on these results alone, managers would be directed towards this hospital in order to investigate why it is suffering from such a relatively large variation in bed occupancy compared to the other hospitals. A useful performance measure for evaluating hospitals should account for hospital size since this relates to the expected variability. This is where the analytical model can aid hospital analysis by producing results which account for known sources of variation, namely Poisson arrivals and emergency patients' lengths of stay. We demonstrate how the modelled results being used as performance indicators can give a much more informative description of hospital performance.

4.4.2 Evaluating hospital performance using model-based performance indicators

Having emphasised the limitations of analysing observed results alone and the importance of using the analytical model for evaluating hospital performance, we now extend our analysis using observed and modelled results together in order to provide a more informed interpretation of the hospital performance. This subsection begins by comparing the observed and modelled results for the CoV, percentage of busy days, and peak days where we identify similarities between both sets of results as well as examine cases where the observed and modelled results do not match. Then we refer back to the earlier analysis based on the observed results alone, except now we introduce the modelled results to show the reader how this improves our understanding of the hospitals.

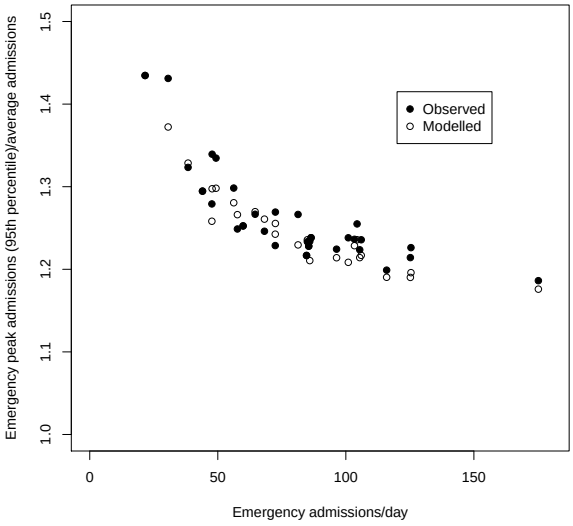
Figure 4.2 shows the observed and modelled results for the arrivals CoV, percentage of arrival busy days and peak arrivals. In all three plots (Figures 4.2a, 4.2b, and 4.2c) the observed results (solid circles) closely match the corresponding modelled results (open circles) indicating the main factors influencing the vari-

ability among arrivals have been captured by the analytical model i.e. the Poisson nature of emergency arrivals. Given the limited control hospitals have over emergency arrivals (which are not planned or typically turned away) we would expect the model to provide a good reflection of the variability of arrivals.



(a) Arrivals coefficient of variation

(b) Arrivals busy days



(c) Arrivals peak days

Figure 4.2: Variation in the emergency admissions per day by mean number of admissions.

Figure 4.2a shows a decreasing trend in the CoV of arrivals as the number of admissions per day increases in both the observed and modelled results. This

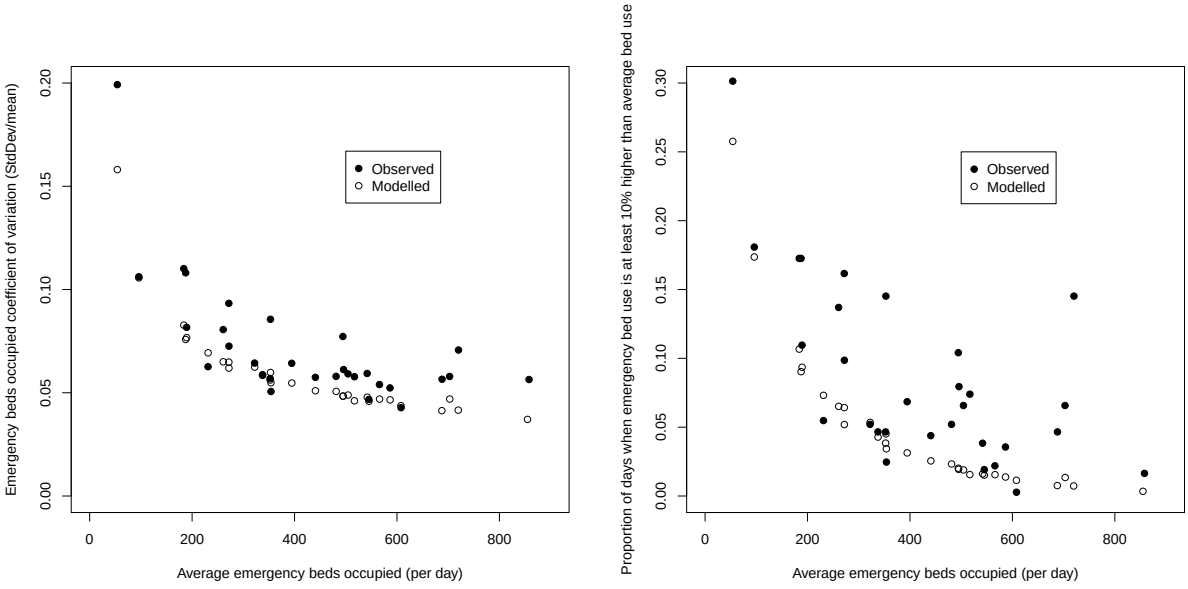
suggests the relative variability among arrivals decreases as hospital size (measured by admissions per day) increases. The same decreasing pattern found in the CoV of arrivals is shown in Figures 4.2b and 4.2c for the busy days and ratio of peak arrivals respectively.

The decreasing trend in all three performance measures for arrivals is greatest among the smallest hospitals (50 or fewer admissions per day). The results continue to decrease for larger hospitals (greater than 50 admissions per day) but the change is less marked with Figures 4.2a, 4.2b and 4.2c showing a more gradual decrease for larger hospitals in comparison to the steeper decline in the smaller hospitals.

The occupancy results indicate a similar story to the arrivals; all three measures of variability have a decreasing trend as hospital size (measured by the average number of beds occupied per day) increases (Figures 4.3a, 4.3b, and 4.3c). This is similar to what we saw with the arrivals, smaller hospitals suffer from relatively large variability in emergency bed occupancy in comparison to larger hospitals.

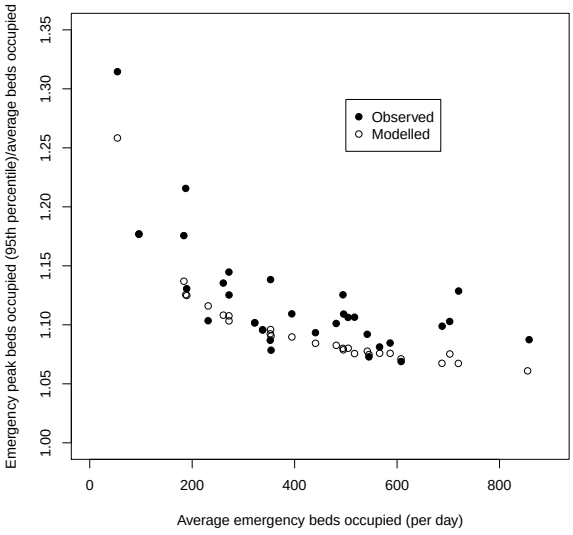
However in comparison to the modelled results for the busy days the observed results are far more variable (Figure 4.3b). For example the 2^{nd} and 4^{th} largest hospitals (an average of 720 and 688 emergency beds occupied per day respectively) are similar in size and share similar modelled results for the busy days (0.0073 and 0.0076 respectively). However the observed busy days for the 2^{nd} largest hospital, 0.145, is almost three times larger than the busy days for the 4^{th} largest hospital, 0.047.

This result could be due to different practices in each hospital's management of bed resources. Consider Figure 4.3c, the observed occupancy peak days are much less variable than their counterpart busy days results. We would expect hospitals to be more concerned when occupancy approaches capacity as opposed to during periods where the occupancy may be higher than average but not necessarily close to breaching capacity. This is because when the demand exceeds the capacity i.e. when there is a bed blockage, this can have a knock on effect and cause further problems such as increased patient waiting times, therefore hospitals



(a) Occupancy coefficient of variation

(b) Occupancy busy days



(c) Occupancy peak days

Figure 4.3: Variation in the emergency occupancy per day by hospital size.

make a greater effort to avoid days where the occupancy nears capacity. Whereas (some) hospitals may be prepared to tolerate more busy days i.e. days of increased pressure on the system, but where the occupancy is not causing the hospital any undue strain.

The occupancy plots (Figures 4.3a, 4.3b, and 4.3c) show the modelled results sitting below their related observed results for the majority of hospitals for all three

measures of variability. Since the performance indicators provide the expected level of variability for each hospital given the pattern of emergency arrivals and patient length of stay distributions and all other things being equal, this indicates the majority of hospitals are contending with additional sources of variability not incorporated into the analytical model. We later discuss how the modelled results could be described as a 'lower-bound', explaining why for the majority of cases they are smaller than the observed results.

In Subsection 4.4.1 we conducted an analysis using traditional performance indicators and identified the four worst performing hospitals based on the ratio of the observed over the mean beds occupied CoV. We also pointed out this approach does not take into account the expected variability at each hospital and subsequently these results for assessing hospital performance are clouded by hospital size. We follow this up but now employ the modelled results to remove size as a factor when investigating the performance of the hospitals.

Table 4.2 presents two sets of results; the ratio of observed over the mean CoV (used in the earlier analysis), and the observed over the modelled CoV - our new criteria for ranking hospitals. The ranks of each hospital under each set of results are included so that we may compare how incorporating the modelled results has improved our ability to interpret the performance of the hospitals. For example the worst performing hospital based on the criteria using the observed results only, Hospital 24, is now ranked 10^{th} according to our new results. What was previously an outlier in Figure 4.1, providing a strong indication this hospital was performing worse than the others, is now relatively unremarkable (in comparison to the other hospitals) and does not necessarily warrant closer inspection. By using the model we are able to remove hospital size as a factor when measuring the hospital performance and with Hospital 24 being the smallest, according the the average beds occupied per day, it is no surprise its observed over the mean result was much smaller compared to the other hospitals.

Previously hospital size was masking their true performance, where the hos-

Hospital number	Rank of observed beds CoV	Ratio of observed/mean beds CoV	Rank of observed/modelled beds CoV ratio	Ratio of observed/modelled beds CoV
24	1	2.76	10	1.26
19	2	1.53	8	1.33
23	3	1.50	6	1.43
25	4	1.47	27	1.00
21	5	1.29	4	1.44
16	6	1.19	5	1.43
30	7	1.13	22	1.06
18	8	1.12	13	1.24
28	9	1.07	2	1.60
22	10	1.01	17	1.17
26	11	0.98	1	1.70
29	12	0.89	23	1.03
10	13	0.89	16	1.18
20	14	0.87	30	0.90
17	15	0.85	9	1.27
4	16	0.82	12	1.24
6	17	0.82	15	1.21
15	18	0.81	26	1.01
27	19	0.80	19	1.14
1	20	0.80	14	1.23
8	21	0.80	11	1.25
12	22	0.80	20	1.13
14	23	0.79	25	1.01
2	24	0.78	7	1.37
7	25	0.78	3	1.52
3	26	0.75	18	1.15
5	27	0.73	21	1.12
13	28	0.70	29	0.92
9	29	0.65	24	1.02
11	30	0.59	28	0.98

Table 4.2: Table showing the observed beds occupied coefficient of variation and the ratio of observed to modelled beds occupied coefficient of variation with their rank by size in descending order.

pitals which were reported as performing worst from the analysis based on the observed results only (Figure 4.1) were the smallest hospitals. The larger hospitals were never likely to be highlighted for underperforming when using the old criteria due larger hospitals experiencing relatively small variability in occupancy and size not being incorporated into the performance measure. However this is no longer the case with Hospital 26, one of the largest hospitals, exhibiting the

greatest observed over modelled CoV (1.70) whereas previously it was only the 11^{th} largest for the observed over the mean CoV (0.981). Given the expected variability, Hospital 26 appears to be struggling with a relatively large CoV.

The histogram of the ratio of the observed over the modelled CoV, Figure 4.4, does not show an extreme case like we detected in the analysis using the observed over the mean CoV (Figure 4.1). In fact there are fewer cases which are "outliers", with the majority of results more narrowly spread (between 1.0 and 1.5) compared to the previous histogram, Figure 4.1. This indicates that most hospitals seem to be coping with a similar amount of variability in occupancy relative to their size.

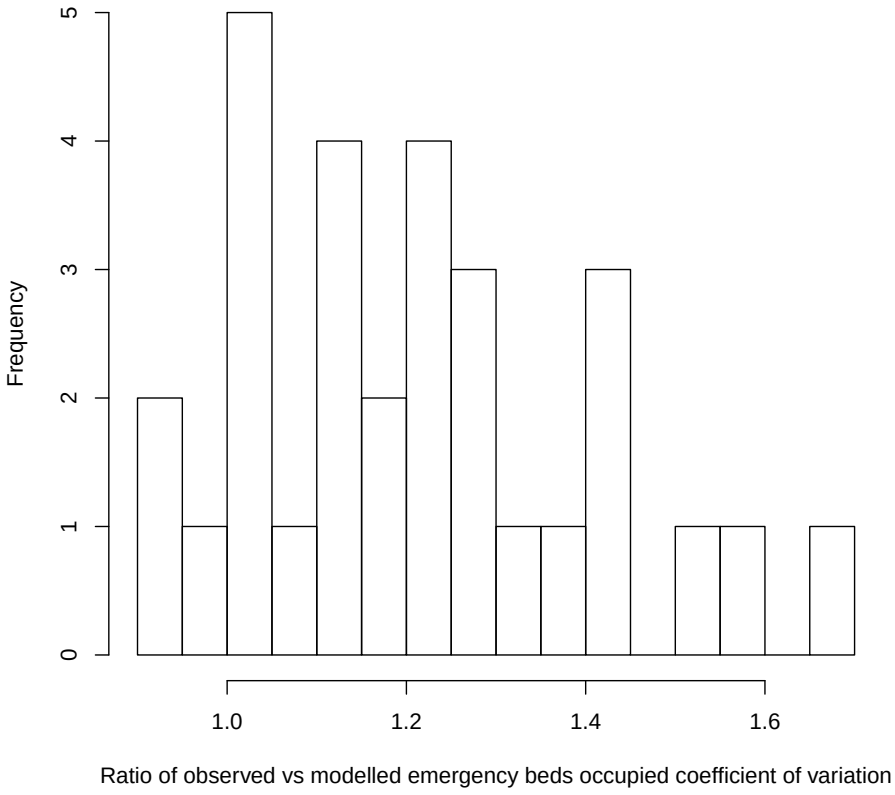


Figure 4.4: Histogram of the ratio of the observed over the modelled beds occupied coefficient of variation.

4.5 Discussion

Evaluating the performance of hospitals is not straightforward given that hospitals are subject to different levels of variability which needs to be incorporated into any results describing their performance. With this in mind we set out to demonstrate the use of our analytical model, factoring in known sources of variability, namely random emergency arrivals, day-of-week effect in arrivals and patients' lengths of stay distributions, to produce modelled results for hospitals dealing with emergency patients.

Modelling can help decision makers and hospital managers understand what to expect in terms of the level of variability faced by hospitals. Studying the modelled results alone provides an indication of the variability of bed occupancy a hospital should expect. Furthermore, using the modelled results as a point of comparison for the observed results, i.e. developing model-based performance indicators, helps provide a more accurate description of hospital performance by accounting for expected variability.

The model does have its limitations and we should be wary of using the performance indicators as a simple tool for judging whether or not a hospital is being managed poorly. Instead a more constructive use of the performance indicators would be to draw attention to hospitals which (potentially) require further investigation, to determine why that level of variability is being exhibited. This not only includes hospitals with large variability, relative to their respective performance indicators, but also the (apparently) "best" performing hospitals, those with variability at or even below the modelled results. Further investigation could then identify characteristics of the best and worst performing hospitals and shed light on whether this is due to good management practices or if there are other (justifiable) factors impacting on the variability of emergency bed occupancy in hospitals.

This perspective of performance indicators, as a tool for highlighting and instigating further investigation reflects the difficulty of evaluating and comparing

hospital performance. For example it would be unfair to directly compare a small hospital with a large hospital due to the relatively large variability subject to the smaller hospital. In practical terms this means we expect smaller hospitals have to prepare for relatively large peaks and troughs in emergency bed occupancy resulting in a need for a greater reserve capacity in beds (and other resources including staffing levels) above the average workload to accommodate the increased variability. On the other hand larger hospitals should be better equipped to cope with the relatively small variability in emergency bed occupancy.

We noted the modelled occupancy results were lower than the observed results in general. One way of interpreting this could be the modelled variability provides a 'lower-bound' or 'best case' description of the variation in the bed occupancy. Emergency work is subject to random arrivals which also vary by day of week and patients' lengths of stay which the performance indicators account for. But possible causes of increased variation in the actual occupancy data could include systematic patterns to discharges (such as fewer discharges at weekends), seasonality in arrivals or length of stay, or 'bull-whip' effects where a system over-compensates in response to fluctuations in inputs (e.g. when occupancy is nearing capacity causes an increase in discharges to release pressure on bed demand, leading a lowered occupancy level and fewer discharges).

We have not considered the impact of ward levels in our model. In this chapter we have summarised hospitals as a whole focusing on the variability in the overall hospital occupancy which is simpler than considering individual wards and provides some useful insights into hospital performance. However we cannot assume that focusing on the total hospital bed stock is sufficient to fully understand the variability in bed occupancy, as this does not necessarily mean a bed on a particular ward is available. Individual wards could be characterised by the analytical model in the same way entire hospitals have been modelled here, given ward arrival and occupancy data. Users may decide this level of modelling is warranted for critical wards where placing patients off-ward is particularly undesirable e.g.

placing patients needing to attend an intensive care unit (ICU) off-ward is best avoided where possible.

We have focused entirely on emergency work in this chapter but have not considered the possible impact of elective patient demand. Since emergency arrivals occur randomly and cannot be turned away the hospital has limited control over the emergency patient workload however elective patients arrive according to some admission schedule. Therefore a greater level of control can be expected over elective patients. In well managed hospitals we would expect to observe a negative correlation between the emergency and elective workloads; as emergency bed occupancy increases the elective occupancy should decrease in order to ease the demand on beds within the hospital, and as emergency occupancy decreases the elective occupancy should increase in order to maximise patient throughput and not waste available resources (beds). Hospitals may employ a more complex system than we have described here. For example, certain wards may be dedicated to selected elective patient types while other wards are more flexible accommodating emergency patients as and when required. It may be of interest for hospitals to analyse their emergency work alongside the elective work - are smaller hospitals better organised in balancing the two workloads or does the relatively large variability cause problems making efficient management of hospital beds difficult?

This chapter has introduced model-based performance indicators as a tool for evaluating and comparing the emergency work of 30 hospitals of varying size. We have shown how the performance indicators factor in sources of variability to provide a performance measure with which to evaluate hospitals against regardless of size. We have highlighted the important role performance indicators can play when evaluating hospital performance by comparing a method of ranking hospitals with and without the use of the modelled results. This demonstrated how our model-based performance indicator, the ratio of observed over modelled CoV, provides a more informed description of hospital performance which would not normally be available to hospital managers without the model. There are limita-

tions to the performance indicators, for example other sources of variability known to the hospital may explain the level of variability in some cases, which is why we argue that the model-based performance indicators should be used as a tool for gaining insight into hospital performance and encourage further investigation into individual cases.

Chapter 5

Model-based performance indicators for elective work

5.1 Introduction

The purpose of this chapter is to develop and evaluate model-based performance indicators for elective and total workloads (combining elective and emergency workloads). We use the performance indicators in the same fashion as we did for the emergency work, to measure hospital performance accounting for known sources of variability and to encourage further investigation into the best and worst performing hospitals.

We present the elective work of the same 30 hospitals in Chapter 4 where we focused on the counterpart emergency work. The same basic approach used for the emergency workload is applied to the elective workload, however in doing this a number of issues arose that were not encountered when investigating the emergency workload:

- arrivals are not Poisson, instead we characterise elective arrivals with a general arrival distribution,
- variations in occupancy levels between weekdays and weekends are deliberate,

- evidence of day-of-week dependent patient lengths of stay,
- day-of-week dependent length of stay parameters may not be directly available from the data but must be estimated.

The analytical model we use to characterise the elective patient demand is the same one presented in Section 3.2. We briefly remind the reader of the main features of the model including how we have accommodated elective patient arrivals.

A key difference between modelling the elective work as opposed to the emergency work is the incorporation of day-of-week dependent length of stay distributions whereas previously we had focused on a single length of stay distribution. Given restrictions on the data it is necessary to use a calibration process to estimate the day-of-week dependent length of stay parameters, though this method is subject to sampling variation. Calibrating the model presents an important but unexpected challenge which requires further investigation to determine how factors such as the number of arrivals, length of stay parameter values and data size impact on the parameter estimation. In this chapter we present the modelled results using estimated parameters from the calibration process while a statistical analysis of the calibration process to determine the impact of sampling error on our parameter estimation is the subject of Chapter 6.

An important part of this chapter is to extend our analysis to both elective and emergency work together. The total modelled bed occupancy is defined as the sum of the elective and emergency bed occupancy. While we model these two streams of work separately and evaluate hospitals based on each of these patient types, elective and emergency patients may not be independent within a hospital. Typically hospital managers will attempt to balance elective and emergency work, smoothing the occupancy as best they can while avoiding bed blockages or patient overcrowding. In well run hospitals we might expect a negative correlation between elective and emergency bed occupancy. During periods of high emergency bed demand hospital managers may respond by reducing the incoming elective demand, possibly by rescheduling or cancelling elective admissions, and vice versa during

low emergency demand managers will take advantage and maximise elective patient throughput. For example if a hospital experiences low emergency admissions over the weekend it may well anticipate low emergency demand on Monday and try to maximise elective patient throughput correspondingly. However balancing elective and emergency work is not straightforward, if the bed occupancy of just one type of work is particularly variable within a hospital, this can cause difficulties managing the hospital as a whole. This concept of good hospital management is contestable, one could easily argue that maintaining elective patient throughput despite variability in emergency patient demand is also indicative of a well run hospital. There is also the argument that when managers have limited (if any) notice of low emergency demand, are able to schedule additional elective work? In these cases not having negative correlation between elective and emergency bed occupancy is not necessarily a symptom of poor management.

The aim of this chapter is to investigate modelling elective patient demand with our analytical model and produce model-based performance indicators for 30 hospitals. We analyse the elective work of hospitals and discuss the dangers of focusing on the elective performance indicators alone, and explain that they should be considered alongside the emergency and total results to avoid misinterpreting the performance indicators. We demonstrate using the performance indicators to evaluate the elective and total workloads of the hospitals, ranking their performance and indicating which cases potentially need further investigation.

This chapter begins with the initial modelling process of the elective work using a single length of stay distribution, analysing the results and consequent decisions to refine our methodology, including focusing on weekday patient demand only (Section 5.2). This is followed by a description of the model under new assumptions using day-of-week dependent length of stay distributions which required a calibration process to estimate the length of stay parameters (Section 5.3). The observed and modelled results for the chosen performance measures are presented, where we discuss the results using the performance indicators to evaluate the

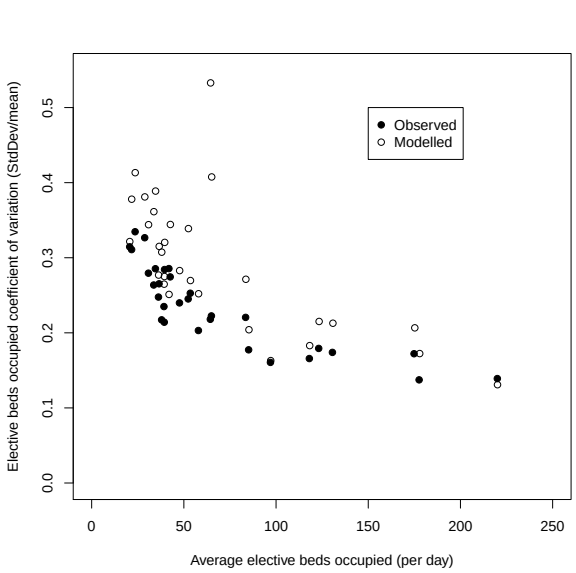
hospitals (Section 5.4). We then extend this work to include emergency patient demand as well, analysing results on the total occupancy at hospitals (Section 5.5). Finally, in Section 5.6, we discuss both the use of performance indicators to evaluate hospital performance for elective work and how our analysis of elective and emergency work together can enhance our understanding of the hospitals.

5.2 Modelling elective work using a single length of stay distribution

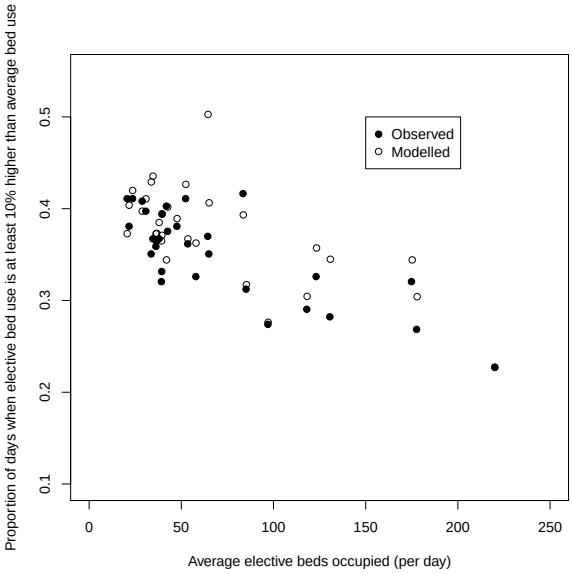
The same analytical model used for the emergency work is applied here but using a general arrival distribution. We assumed emergency arrivals follow a Poisson distribution to mirror their random, unexpected nature however for elective patients a general arrival distribution is more appropriate because we assume managers are able to exert some degree of control in the form of a planned elective admission schedule. We bear in mind that elective arrivals are not entirely deterministic, there remains some uncertainty since patients may turn up late or even not arrive at all. The length of stay distribution remains the same as in the emergency work, a geometric distribution.

The early stages of the elective work mirror the emergency modelling, plotting hospital size against the same three performance measures; coefficient of variation (CoV), percentage of busy days, and the peak days (Figure 5.1). All three plots reveal the modelled results tend to be greater than the observed, suggesting our model is overestimating the variability. Furthermore the beds occupied busy days (Figure 5.1b) show a less pronounced decreasing trend with hospital size than what we had observed with the emergency work, though we note the relatively smaller hospital size for the elective work.

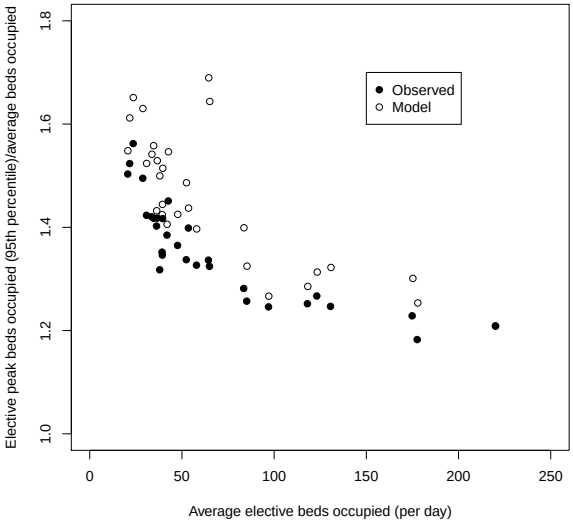
We refine our modelling process by focusing on weekday variability only, reasoning that hospitals deliberately plan elective admissions to reduce the occupancy at weekends while maximising the bed occupancy during the week. Typically hos-



(a) Elective beds coefficient of variation



(b) Elective beds busy days



(c) Elective beds peak days

Figure 5.1: Variation in the elective beds occupied per day by hospital size.

pitals lighten the bed demand at weekends due to having fewer staff during these days. In some sense hospitals are deliberately inducing variability between weekday and weekend occupancy but this is not the type of variability we want to identify with our model to measure hospital performance as this is not necessarily an indicator of poor hospital management.

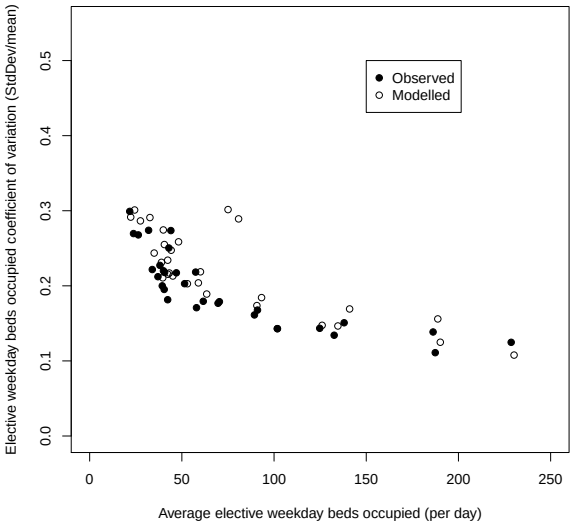
Figure 5.2 shows our updated performance measures after removing the week-

end data, where we can see the observed and modelled results matching more closely. The percentage of busy days in Figure 5.2b now shows a strong decreasing trend as hospital size increases, also the range of the percentage of busy days has dropped from approximately (0.2, 0.5) to (0.1, 0.45) in the weekday only case. These factors indicate removing weekends from our elective analysis has improved our modelling, however the modelled results are still greater than the observed results in the main. Given that we have previously argued the model provides a best-case scenario for hospitals and would produce results below the observed, it seems likely that there remain other factors in the elective workload which we have yet to incorporate into our model such as day-of-week dependent patient lengths of stay.

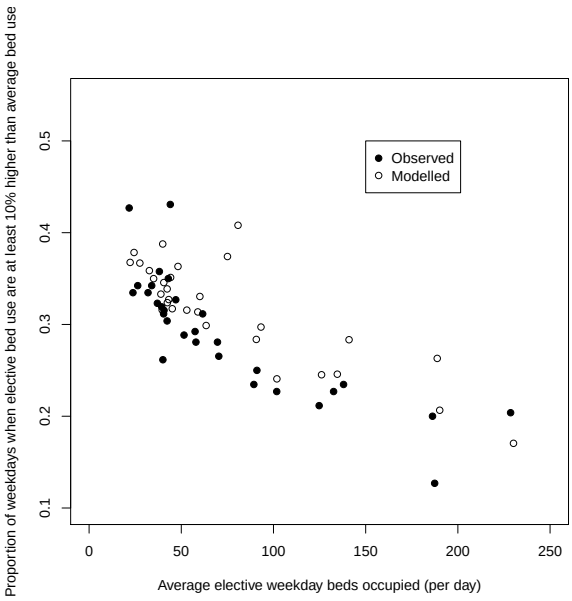
5.3 Modelling elective work with day-of-week dependent length of stay distributions

Hospitals may employ an elective scheduling plan which aims to admit patients of the same type on the same day. A reason for this could be the hospital trying to smooth the bed occupancy during the week, patient types with long lengths of stay are booked earlier in the week, while those with shorter expected lengths of stay are admitted towards the end of the week leading to a smoother bed census during the week and the majority of patients being discharged by the weekend.

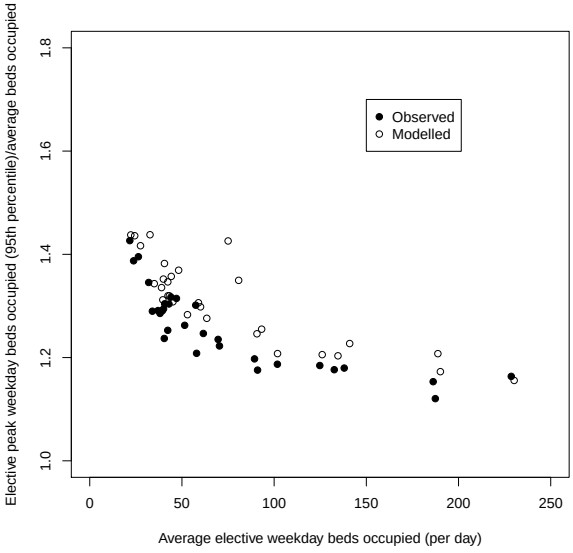
To improve our modelling of elective patient demand we incorporate day-of-week dependent patient length of stay distributions which requires estimating the length of stay parameters for each day of the week. However given that the available data does not provide lengths of stay by day of arrival, this is not straightforward. Our approach to calibrate the model estimates the length of stay parameters as those which minimise the difference between the observed occupancy (provided in the data) and the modelled occupancy. It should be noted that any statistical calibration process of this sort is subject to random variation as they are based



(a) Elective weekday beds coefficient of variation



(b) Elective weekday beds busy days



(c) Elective weekday beds peak days

Figure 5.2: Variation in the elective beds occupied per day by hospital size excluding weekend data.

on a finite sample of patient data which affects the parameter estimation. A detailed description and a statistical analysis of the calibration process is included in Chapter 6 of the thesis. For the purposes of this chapter we apply the parameter estimates as computed by the calibration process.

Figure 5.3 shows an improved fit between the modelled results (using day-of-week dependent length of stay distributions) and the observed results. There are

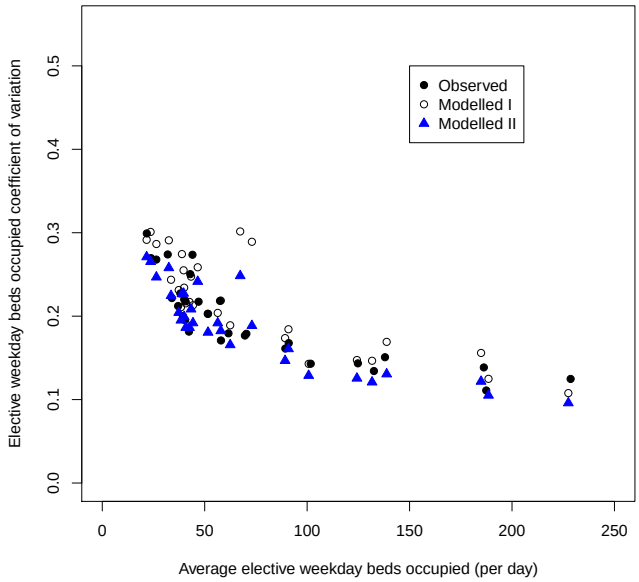


Figure 5.3: Elective CoV with modelled results using single (Modelled I) and day-of-week dependent (Modelled II) length of stay distributions excluding weekend data.

fewer outliers compared to the single length of stay distribution modelled results and the remaining outliers are less pronounced. We have focused on the beds occupied CoV to demonstrate the impact of incorporating day-of-week dependent length of stay distributions though the results for the percentage of busy days and the peak occupancy show a similar story.

Introducing day-of-week dependent length of stay distributions has reduced the variability of the modelled occupancy compared to using a single length of stay distribution. This effect is shown in Figure 5.3 where for every hospital the modelled CoV using a single length of stay distribution (Modelled I) is greater than the day-of-week dependent results (Modelled II).

5.4 Evaluating elective hospital performance using model-based performance indicators

In this section we present a full implementation of our model to compute performance indicators for each hospital, excluding weekends and incorporating day-of-week dependent length of stay distributions, to provide a similar analysis conducted with the emergency work. However as a result of the calibration process we bear in mind the elective hospital results are subject to random variation, hence one should be wary of over-interpreting difference in results between hospitals, especially those that are close to one another.

We must be wary of over-interpreting the results for the elective workload alone. Using the variability of occupancy as a way of measuring hospital performance assumes hospitals seek to avoid large peaks and troughs in bed occupancy. We can still view relatively small observed variability as an indicator of good hospital management. However with elective work, there are cases where hospitals may exhibit large variability but are being managed well. The priority of a hospital is not necessarily focused on elective work alone but rather smoothing the total occupancy, which may mean the variability of the elective work is relatively large in response to the variability of emergency work. In this case the elective variability would suffer but the total hospital occupancy would be performing well as a result. With this in mind we first present the results of the elective work similar to the emergency work but followed by an analysis of the total bed occupancy results.

Figure 5.4 and Table 5.1 shows the modelled results, using day-of-week dependent length of stay distributions, are mainly lower than the observed in the majority of cases. Since the performance indicators represent the expected variability at each hospital this indicates the hospitals are suffering from additional sources of variability not included in the analytical model.

There remain some cases where the observed results are smaller than the modelled results, for example Hospital 27 (which has an average of 70 beds occupied per

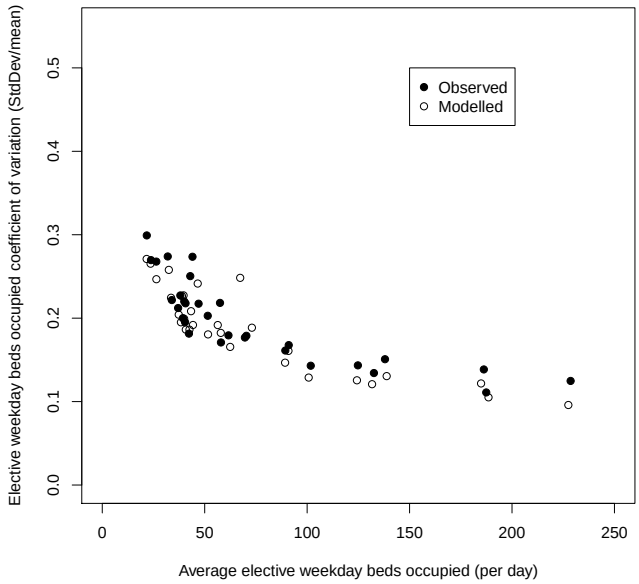


Figure 5.4: Elective CoV with day-of-week dependent length of stay distributions excluding weekend data.

weekday) has an observed result (0.179) smaller than the modelled (0.248) (Table 5.1). This arguably indicates the hospital is performing well somehow managing to keep their elective occupancy relatively low compared to the expected variability. This could possibly be due to the hospital monitoring their occupancy and if during the week the hospital is coming under pressure as patient demand nears capacity, they react by lowering their elective admissions. This could mean cancelling elective patient arrivals (which is not an option with emergency work) in order to reduce the peak in demand and thus smoothing the bed census, reducing the variability in occupancy.

Whilst hospital size is an important factor in predicting variability in occupancy, it is not the only factor. For example hospitals 27 and 4, the 10^{th} and 12^{th} largest hospitals (70 and 62 average beds occupied per weekday respectively), are similar in size but exhibit very different modelled results (0.248 and 0.166 CoV respectively, see Table 5.1). This indicates hospital size is not the sole driver of expected variability in bed occupancy but other factors incorporated into the model such as patient arrivals and lengths of stay can have a large impact.

Table 5.1 presents the ratio of observed over the modelled CoV (elective CoV ratio) for each hospital along with their rank based on this statistic. We earlier identified Hospital 27 as having a relatively large modelled CoV result based on its size, Table 5.1 indicates this is the best hospital in terms of the elective CoV ratio, while Hospital 4 is the 15^{th} worst performing hospital - although we are aware hospitals may deliberately induce elective bed variability to balance out emergency work.

The histogram of the elective CoV ratio, Figure 5.5, shows that majority of hospitals have a CoV ratio clustered around 1. The observed results match the modelled results quite closely for most hospitals, indicating that hospitals are exhibiting a level of variability which we would expect given their arrivals and patient lengths of stay - and are not necessarily taking extreme measures to alter their elective workload to accommodate emergency patients.

In Figure 5.5 the most extreme case is Hospital 17, with an elective CoV ratio greater than 1.4 i.e. this hospital is suffering from an observed CoV 40% greater than the modelled CoV. This alone is not enough evidence to state unequivocally the hospital is under-performing but it does encourage further investigation to explain why this is happening. At the other end of the spectrum, one hospital (27) has an elective CoV ratio smaller than 0.8. This hospital also warrants further investigation into why the observed variability is so much lower than the expected variability. The performance indicators offer a way of analysing hospitals not just to identify poorly performing cases but also the best hospitals as well.

5.5 Elective and emergency work

So far we have focused on elective and emergency workloads separately however these two patient types are not necessarily independent. Hospitals try to balance the workload of both elective and emergency patients. During periods of high emergency patient demand well managed hospitals may reduce pressure on the system by reducing the number of elective admissions, and vice versa hospitals

Hospital number	Rank of CoV ratio	Average weekday beds occupied per day	Modelled CoV	Observed CoV	CoV ratio
1	5	138	0.131	0.151	1.16
2	8	186	0.122	0.139	1.14
3	13	89	0.147	0.161	1.10
4	15	62	0.166	0.179	1.08
5	18	91	0.161	0.168	1.04
6	7	57	0.192	0.218	1.14
7	6	125	0.126	0.143	1.14
8	2	229	0.096	0.125	1.30
9	11	102	0.129	0.143	1.11
10	10	133	0.121	0.134	1.11
11	25	42	0.186	0.181	0.97
12	28	58	0.182	0.171	0.94
13	27	70	0.189	0.177	0.94
14	26	40	0.227	0.220	0.97
15	29	47	0.242	0.217	0.90
16	23	34	0.224	0.222	0.99
17	1	44	0.192	0.274	1.43
18	24	40	0.199	0.195	0.98
19	20	39	0.195	0.200	1.02
20	22	38	0.227	0.227	1.00
21	16	32	0.258	0.274	1.06
22	12	22	0.271	0.299	1.10
23	4	41	0.186	0.218	1.17
24	19	37	0.204	0.212	1.04
25	21	24	0.265	0.270	1.02
26	17	187	0.105	0.111	1.06
27	30	70	0.248	0.179	0.72
28	9	51	0.181	0.203	1.12
29	3	43	0.208	0.250	1.20
30	14	26	0.247	0.268	1.09

Table 5.1: Table listing the elective weekday CoV observed and modelled results.

may attempt to admit more elective patients during quiet periods. Since elective and emergency patients can influence hospital management it is important to analyse them together.

Assuming independence of elective and emergency work we compute the mean and variance of the total (weekday) bed occupancy by summing the contributions from the elective and emergency work. We note that independence is only nec-

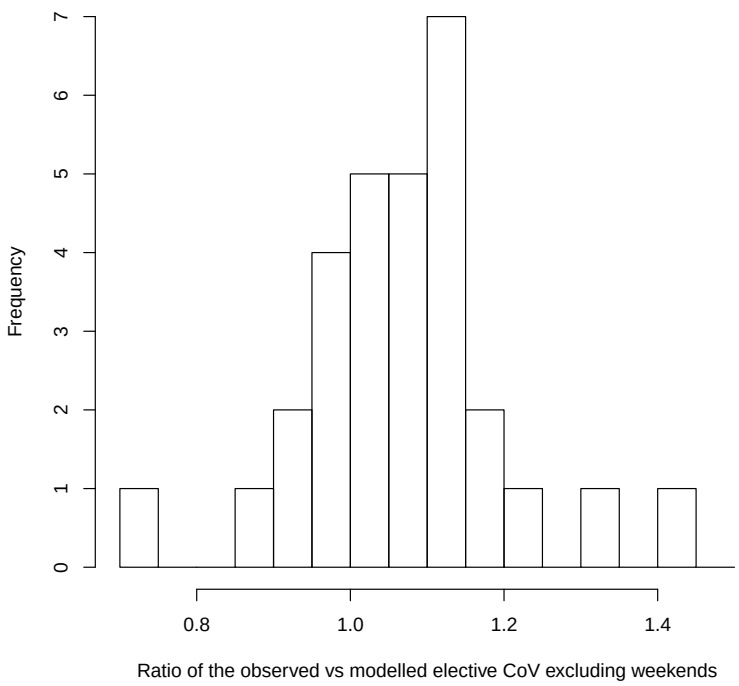
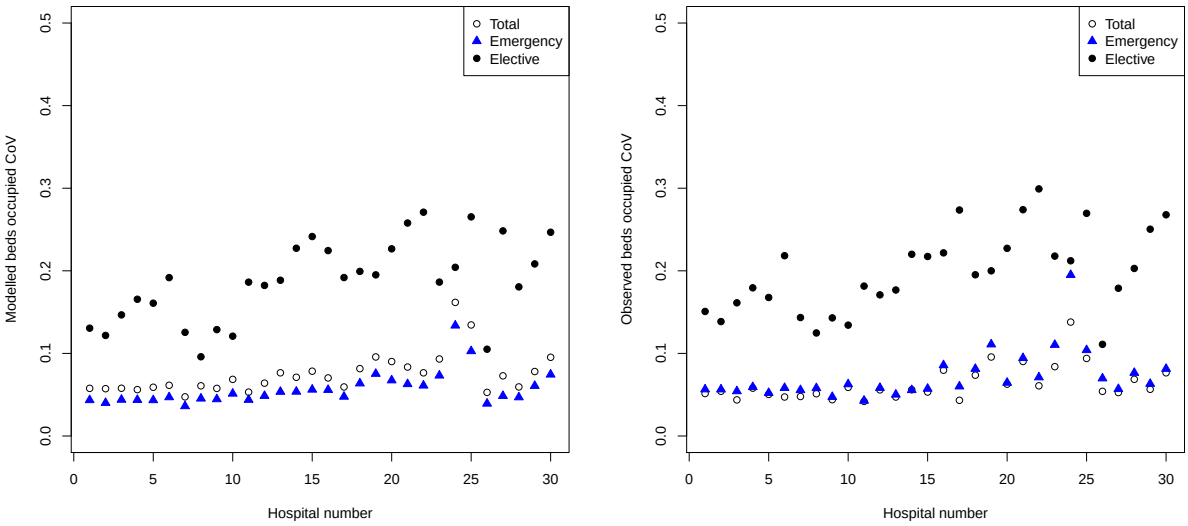


Figure 5.5: Histogram of the elective beds occupied CoV ratio excluding weekends.

essary to calculate the variance of total bed occupancy and not the mean. With the mean and variance of total bed occupancy we can calculate the performance measures for the total occupancy.

When analysing the emergency work in Chapter 4 we used data on the entire week. However to remain consistent with the weekday analysis conducted with the elective work we focus only on the weekday demand for both elective and emergency work here.

If elective and emergency patients were treated independently, as the model assumes, then the total CoV is greater than the emergency CoV (Figure 5.6a), whereas the opposite is true for the observed results (Figure 5.6b) with the total CoV being less than the emergency CoV. This is a consequence of the reality that hospitals do not treat elective and emergency patients independently but try to balance both patient types to smooth the total occupancy. This does not necessarily mean all of the hospitals are managing their total workloads well - they could be suffering from undue variability of one or both patient types. However it



(a) Elective, emergency and total modelled CoV by hospital number

(b) Elective, emergency and total observed CoV by hospital number

Figure 5.6: Comparison of elective, emergency and total beds occupied CoV excluding weekends.

means that some action is being taken towards managing total bed occupancy.

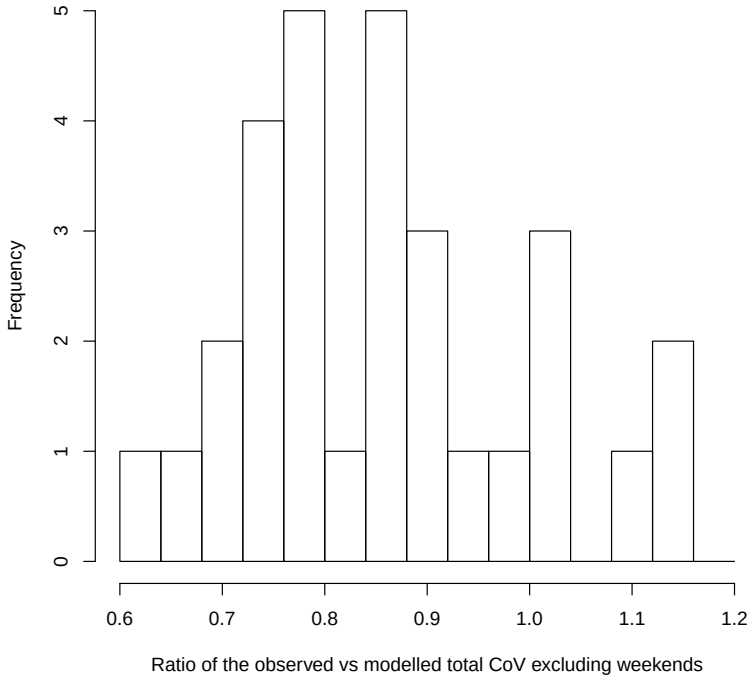


Figure 5.7: Histogram of the total beds occupied CoV ratio excluding weekends.

The histogram of the total CoV ratio (observed over modelled CoV for the total weekday bed occupancy), Figure 5.7, is different from what we observed with the elective and emergency work. For 24 out of the 30 hospitals the total CoV ratio is less than one i.e. the observed CoV is less than its respective modelled CoV. Thus the majority of hospitals are exhibiting less variability in their total bed occupancy compared to the variability if elective and emergency workloads were treated independently.

With the elective and emergency work there had been a pattern of the observed variability being greater than the modelled but for the total bed occupancy only a minority of hospitals fall into this category. Figure 5.7 shows that six hospitals exhibit total CoV ratios greater than 1, with Hospitals 28 and 16 (see Table 5.2) being the worst performing according to this performance indicator with CoV ratios greater than 1.1. In these two cases while hospital management is having an impact in reducing the variability of total bed occupancy (Figure 5.6) the observed CoV is still 10% greater than the modelled.

We can also compare hospitals total CoV ratios with the respective elective and emergency CoV ratios. Hospital 8's ratios suggest management is having a positive impact on their bed occupancy. The elective and emergency ratios which are both greater than 1, 2^{nd} and 12^{th} largest respectively, would not initially indicate this is the case. However the total CoV ratio is relatively small, 0.842 (16^{th} largest). This seems to be an example of a hospital suffering from relatively large variability in emergency bed occupancy but responding with the elective workload, smoothing the total bed occupancy.

We earlier mentioned Hospital 16 as one of the worst performers according to the total CoV ratio. This remains consistent with the emergency CoV ratio which was the 4^{th} largest of the hospitals. However with the elective CoV ratio it performed reasonably well with an elective CoV ratio less than 1 (23^{rd} largest elective CoV ratio, Table 5.1). This illustrates the danger of focusing on only the elective performance indicators, without considering the impact on the total

Hospital number	Average total weekday beds occupied per day	Total observed CoV	Total modelled CoV	Rank of CoV ratio	Total CoV ratio	Rank of emergency CoV ratio	Emergency CoV ratio
1	839	0.051	0.058	11	0.892	11	1.299
2	874	0.054	0.057	8	0.945	9	1.408
3	654	0.044	0.058	23	0.756	16	1.234
4	604	0.058	0.056	4	1.032	10	1.356
5	677	0.050	0.059	14	0.854	18	1.201
6	562	0.047	0.061	21	0.769	15	1.239
7	983	0.048	0.047	6	1.009	3	1.539
8	745	0.051	0.061	16	0.842	12	1.278
9	647	0.044	0.058	22	0.763	23	1.062
10	529	0.059	0.069	13	0.859	17	1.222
11	648	0.042	0.053	19	0.789	28	0.975
12	499	0.056	0.064	12	0.873	19	1.199
13	424	0.047	0.077	30	0.617	30	0.936
14	391	0.056	0.071	20	0.788	25	1.036
15	384	0.053	0.078	29	0.679	26	1.019
16	387	0.080	0.070	2	1.133	4	1.535
17	539	0.043	0.059	24	0.728	14	1.261
18	300	0.074	0.082	9	0.903	13	1.277
19	223	0.096	0.096	7	0.999	7	1.472
20	269	0.063	0.090	28	0.697	29	0.957
21	303	0.090	0.084	3	1.081	6	1.503
22	293	0.061	0.077	18	0.794	21	1.163
23	228	0.084	0.093	10	0.900	5	1.508
24	91	0.138	0.162	15	0.852	8	1.458
25	120	0.094	0.134	27	0.698	27	1.012
26	907	0.054	0.053	5	1.024	1	1.779
27	551	0.053	0.073	26	0.722	20	1.172
28	545	0.069	0.059	1	1.156	2	1.632
29	364	0.057	0.078	25	0.725	24	1.041
30	215	0.077	0.095	17	0.803	22	1.093

Table 5.2: Table listing the total and emergency weekday beds occupied CoV ratios of observed over modelled results.

occupancy. This hospital may be suffering from large variability in the emergency workload and has not done enough to balance this with their elective workload, hence the large total CoV ratio.

One of the most consistent hospitals according to our performance indicators

is Hospital 13. This hospital has the smallest emergency CoV ratio (0.936), one of three hospitals to manage a result less than 1 for the emergency weekday only work (Table 5.2). With their emergency work being relatively stable we would expect the elective CoV ratio to be relatively small too (since there is less need to increase and decrease elective patient demand to balance the emergency demand). Table 5.1 shows Hospital 13 has an elective CoV ratio of 0.94, the 4^{th} smallest. As a result it is not surprising this hospital has the smallest total CoV ratio (0.617). This type of analysis with the performance indicators would encourage further investigation of this hospital, possibly to reveal good management practices leading to the low variability among its bed occupancy.

With the model assuming independence of elective and emergency patients, the performance indicators highlight that most of the hospitals are managing their total bed occupancy to some extent resulting in a lower CoV. Using the analytical model to produce performance indicators allows us to compare hospital elective and emergency work performances separately as well as the performance of the total occupancy which provides a much better insight into the hospital management than simply focusing on any one performance indicator.

5.6 Discussion

The nature of elective work is more complex than its emergency counterpart in a modelling context for several reasons: arrivals are not random but should follow some planned admission schedule, weekend occupancy is smaller than weekday occupancy and patient lengths of stay may vary by the day of week. This chapter demonstrates that the analytical model is able to accommodate these traits in elective patient work using a general arrival distribution, excluding weekend data and incorporating day-of-week dependent patient length of stay distributions. These features are important for characterising elective patient demand in order to produce meaningful performance indicators.

Variability of bed occupancy has been key in measuring hospital performance,

where typically hospitals being well managed will smooth their occupancy to reduce variability. However hospitals may deliberately reduce their elective demand at weekends compared to the weekdays. We removed this intentional variability from our analysis as this could potentially mask a hospital's true performance.

Incorporating day-of-week dependent length of stay distributions required estimating the average length of stay parameter for each day of the week. However given data restrictions, i.e. not having daily patient arrivals and departures meant this is not straightforward. In the following chapter we provide a statistical analysis of the calibration process and how it is subject to random error.

The elective workloads of the hospitals were smaller than their respective emergency workloads but we found similarities between the two in terms of hospital performances. The performance indicators, which account for hospital size, revealed a decreasing trend as hospital size increased. It is not appropriate to analyse the elective work in the same manner as the emergency work because large variability in emergency work is not necessarily a result of poor hospital management. It may be a response to large variations in emergency bed occupancy, aiming to smooth the total bed occupancy, which is often a main priority for hospitals.

Initially we focused on elective and emergency work separately but hospitals should endeavour to balance elective and emergency work, so reducing the variability of their total occupancy. Hospitals may be willing to accept more variable elective and/or emergency work if that means smoothing the total occupancy census, therefore measuring hospital performance by total bed occupancy provides a more useful indicator of a hospital's true performance.

An important development in analysing hospital performance is combining the elective and emergency work to compute the total bed occupancy. The modelled results for total bed occupancy assume elective and emergency patients were treated independently. Although this may be unrealistic because hospitals should be managing elective and emergency workloads together, the modelled results still provide insights for monitoring hospital performance. Comparing the modelled

and observed total CoV results showed that hospitals are not treating their elective and emergency patients independently but managing them in order to reduce the variability of total occupancy.

Performance indicators for elective, emergency and total bed occupancy allow us to analyse the hospitals and highlight cases where management seems to be performing well or struggling with variability in bed occupancy. Where there are limitations with the performance indicators for the elective workload (due to hospitals possibly allowing variability in their elective work) or the total workload (since the model assumes elective and emergency work are independent) having results for all three categories provides a more informed perspective on each hospital's performance.

We were able to pick out hospitals which were performing poorly according to the total CoV ratio and find some explanation with the elective and emergency CoV ratios. For example a hospital struggling in terms of the total CoV ratio, may not be as a consequence of poorly balancing elective and emergency work.

We can extend this approach of comparing elective and emergency hospital performance by using the performance indicators for total occupancy to rank the hospitals and encourage further investigation into individual cases. The best performing hospitals could be a result of performing well for both elective and emergency work separately, resulting in low variability for total bed occupancy. Alternatively should a hospital be suffering from large peaks and troughs in their emergency bed occupancy, then they may respond by varying their elective work accordingly to smooth the total occupancy. The result would be large variability for both elective and emergency occupancy but low variability for the total occupancy.

This chapter sets out to model elective and total bed occupancy and produce model-based performance indicators. Modelling the elective work differed to the emergency case in that we removed weekend data and incorporated a general arrival distribution and day-of-week dependent length of stay distributions. The latter requiring a calibration process for estimating the length of stay parameters.

Unlike the emergency work, there were limitations in determining hospital performance using the variability of elective work due to hospitals deliberately inducing variability in their elective occupancy to balance their emergency work. For the elective and emergency occupancy the model-based performance indicators offer a best case scenario for the hospitals. However for the total occupancy, underpinning the model is the assumption of independence of elective and emergency patients which we anticipate is not typically the case. Therefore the model-based performance indicators may not reflect precisely the true nature of the elective, emergency, and total workloads but this does not prevent them from being a useful tool for measuring hospital performance and highlighting cases requiring further investigation. By studying the performance indicators of all three categories we gained a more informed insight into hospital performance and discussed why they may be performing well or poorly according to their set of results. This work highlights the role analytical models can play in measuring hospital performance and also the importance of looking at elective and emergency work together.

Chapter 6

Robustness of the model calibration process

6.1 Introduction

Patient lengths of stay are one of the main drivers for hospital occupancy and hence it is important to represent them as accurately as possible in any model-based performance indicators. In some cases it is sufficient to use a single length of stay distribution for a patient type - the approach we took in Chapter 4 for emergency patients where data on the number of arrivals and occupancy per day for a year was provided. In this instance computing the overall average length of stay was relatively straightforward given the data.

In other cases, as noted in Chapter 5 for elective patents, modelling patient workloads may require the use of day-of-week dependent length of stay distributions. Furthermore, as experienced in our analysis of elective workloads, available data may not be sufficient to obtain them directly.

The calibration process used to provide the day-of-week length of stay parameter values using data consisting of the number of arrivals per day and the occupancy per day. The results presented in Chapter 5 are based on the seven estimated parameters (one for each day of the week) derived from the calibra-

tion process. However such methods are always subject to sampling variation. This chapter investigates the main factors affecting the parameter estimation and investigates the accuracy of the calibration process.

In practice the calibration process estimates the day-of-week length of stay parameters by minimising the squared error between the modelled occupancy (where the length of stay parameters are variables) and the occupancy levels. In our investigation we generate simulated data sets consisting of daily occupancies given known arrivals and lengths of stay (which have been sampled from our chosen length of stay distributions). The purpose of the simulated data sets is to allow us to control the true length of stay parameters while producing daily occupancies over the year which are subject to sampling variation. Using the simulated data sets we construct a set of scenarios varying the number of arrivals, length of stay parameter values and data size, where for each scenario we produce 20 sets of parameter estimates. To measure the accuracy of the calibration process we use the standard deviation of the parameter estimates from 20 sets of results.

This chapter investigates the accuracy of the calibration process and explains how we compute the simulated data sets and modelled results (Section 6.2). We present the results of a set of experiments which vary the volume of arrivals, the length of stay parameters values, and the size of the data and analyse the impact of these factors on the accuracy of the calibration process (Sections 6.3, 6.4).

6.2 Experimental design

For any simulated data set the calibration process estimates the length of stay parameters as those which minimise the sum of squared errors between the expected occupancy (derived from the analytical model) and the simulated data set. Mathematically the length of stay parameters are estimated from:

$$\min_{\theta} \sum_{t=42}^{365} (M_t(\boldsymbol{\theta}) - S_t)^2,$$

where

- $M_t(\boldsymbol{\theta}) = \sum_{u=1}^{t} N_u s_{u,t}(\boldsymbol{\theta})$ is the modelled occupancy on day t, given the length of stay parameters $\boldsymbol{\theta}$,
- N_u is the known number of arrivals on day u,
- $s_{u,t}(\boldsymbol{\theta})$ is the probability a patient arrives on day u and survives until at least day t which depends on a set of variables, the length of stay parameters $\boldsymbol{\theta}$, and
- S_t is the occupancy on day t from out simulated data set which depends on the true length of stay parameters and known arrivals.

Note that the expected occupancy on day t, $M_t(\boldsymbol{\theta})$, is a function of the known daily arrivals and the seven day-of-week length of stay parameters, $\boldsymbol{\theta}$. The occupancy on day t from the simulated data set is subject to sampling variation as the patient lengths of stay are randomly sampled from the chosen length of stay distributions. The calibration process compares the expected occupancy, $M_t(\boldsymbol{\theta})$, with the simulated data set, S_t, to find the squared error for all days of the year bar the "warm up period". We sum all of the squared errors to summarise the information into a single statistic forming the objective function.

Since the calibration process is based on the daily occupancy data it is expected that the arrivals and length of stay parameters, the main drivers of occupancy, will affect its accuracy. We anticipate that a larger volume of arrivals will improve the accuracy since changes in the length of stay parameter estimates will have a greater impact on the occupancy. While on the other hand larger length of stay parameter values are expected to worsen the accuracy of the calibration process, with more patients present in the system there are more possible combinations of length of stay parameters making it more difficult for the calibration process to estimate the true parameter values. We also expect more data to improve the accuracy of the calibration process while less data will have the opposite effect.

6.2.1 Deterministic arrivals

Since we are estimating the day-of-week length of stay parameters for elective patients it is sensible to use known arrivals which reflects the nature of elective patients. The deterministic arrivals used in the simulation experiments are therefore based on the (elective) arrival data of Hospital 1 (from the 30 hospitals presented in Chapters 4 and 5) though the arrivals could have been selected from any of the hospitals.

We consider different volumes of patient arrivals using a "baseline" set of arrivals (Hospital 1) for comparison. In our experiments we halve and double the number of arrivals to study the effect this has on the accuracy of the calibration process. In addition we scale the weekend arrivals (Table 6.1); the Saturday and Sunday arrivals are multiplied by 6.5 and 5 respectively to match the average arrivals on each of these days to the average weekday arrivals. This was to study whether changing the mix of volumes of arrivals between days had an effect on the accuracy of the calibration process, in particular increasing the relatively small weekend arrivals to a similar level as the weekday arrivals.

	Arrivals
Baseline	The deterministic arrivals used were Hospital 1's elective arrivals
Halved	All arrivals were halved
Doubled	All arrivals were doubled
Scaled	Saturday and Sunday arrivals were scaled separately such that the average arrivals on each of these days equalled the weekday average arrivals

Table 6.1: Table showing the different patterns of arrivals used in measuring the accuracy of the calibration process.

6.2.2 Length of stay parameters

Larger length of stay parameter values mean patients stay in the system longer and contribute to the occupancy on more days. This may make it more difficult for the calibration process - which is minimising the difference between the simulated occupancy and the modelled occupancy - to estimate the length of stay

parameters since the contribution of patients to the occupancy on any given day is more complex. On the other hand for smaller length of stay parameters, patients contribute to the occupancy on fewer days simplifying the distribution of the occupancy on any given day and hence it is likely to reduce the variability of the parameter estimates.

For example, if all of the length of stay parameters are relatively small i.e. patients stay for fewer days in the system, then the occupancy pattern would be similar to the arrival pattern. As there would be few sets of length of stay parameters which could produce this occupancy pattern, we anticipate the calibration process to be more accurate with fewer possible sets of parameters to consider.

Table 6.2 lists the "baseline" set of length of stay parameter values which represent the average length of stay for each day of the week respectively. Typically elective patients tend to stay in hospital only a few days on average with patients rarely staying longer than a week. In addition the average length of stay on Saturday and Sunday tends to be shorter compared to the rest of the week. Hence the "baseline" set of length of stay parameters are selected with this in mind - the average lengths of stay during the week ranging between 2-4 days while the weekends are shorter. One of our scenarios doubles all the length of stay parameter values for the week to study the impact this has on the accuracy of the calibration process. We also double only Monday's, Tuesday's, and Sunday's length of stay parameters to study whether the variability of the length of stay parameter values has an impact on the calibration process.

	Length of stay parameters						
	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Baseline	2.1	2.5	2.5	3.0	3.5	1.8	1.5
Doubled	4.2	5.0	5.0	6.0	7.0	3.6	3.0
Doubled on Mon, Tue, Sun	4.2	5.0	2.5	3.0	3.5	1.8	3.0

Table 6.2: The length of stay parameters used in measuring the accuracy of the calibration process.

6.2.3 Data size

The size of the data is determined by the number of days of occupancy we include in the calibration process, in our work in Chapter 5 the size of the data is $365-42 = 323$ days. We do not include the first six weeks of occupancies as they are subject to a "warm up period". The occupancies at the beginning of the period depend on the arrivals prior to the start of the data, for example the occupancy on day 1 will only include the arrivals on that day since there have been no previous arrivals (and therefore no possible survivors to day 1). The size of the "warm up period" is chosen based on the probability of patient surviving more than six weeks being very small.

When doubling the data we increase the number of days of occupancy to 646 days (see Table 6.3). The number of days of arrivals (required when computing the simulated and modelled occupancy) are doubled by repeating the baseline arrivals as well as sampling new lengths of stay. Similarly for halving the number of days of occupancy we reduce the number to $323/2 = 161.5$ which we round up to 162 days.

	Data
Baseline	323 days of occupancies - 1 year of data less the six week warm up period
Halved	$323/2 = 161.5$ days which has been rounded up to 162 days of occupancies
Doubled	$323 \times 2 = 646$ days of occupancies

Table 6.3: Table of different data sizes used in measuring the accuracy of the calibration process.

6.3 Results

We conduct 20 runs of the calibration process for all of the different scenarios varying the volume of arrivals, length of stay parameter values and data size to produce length of stay parameter estimates for each day of the week for each run. The results of the calibration process are subject to some sampling error, therefore we are using the 20 runs to detect general patterns and effects of different scenarios.

To measure the accuracy of the calibration process we use the standard deviation of each day's parameter estimates.

6.3.1 Baseline hospital

Included in the set of scenarios we analyse is the "baseline" case which acts as a point of comparison for the accuracy of the calibration process for a typical hospital. This allows us to determine whether changing factors such as the volume of arrivals improved or worsened the accuracy of the calibration process in comparison to our "baseline" case.

The characteristics of the "baseline" hospital were defined earlier (in Section 6.2); the deterministic arrivals were taken from Hospital 1's elective arrivals, a set of length of stay parameters for each day of the week (see Table 6.2), and a data size of 323 days of occupancy.

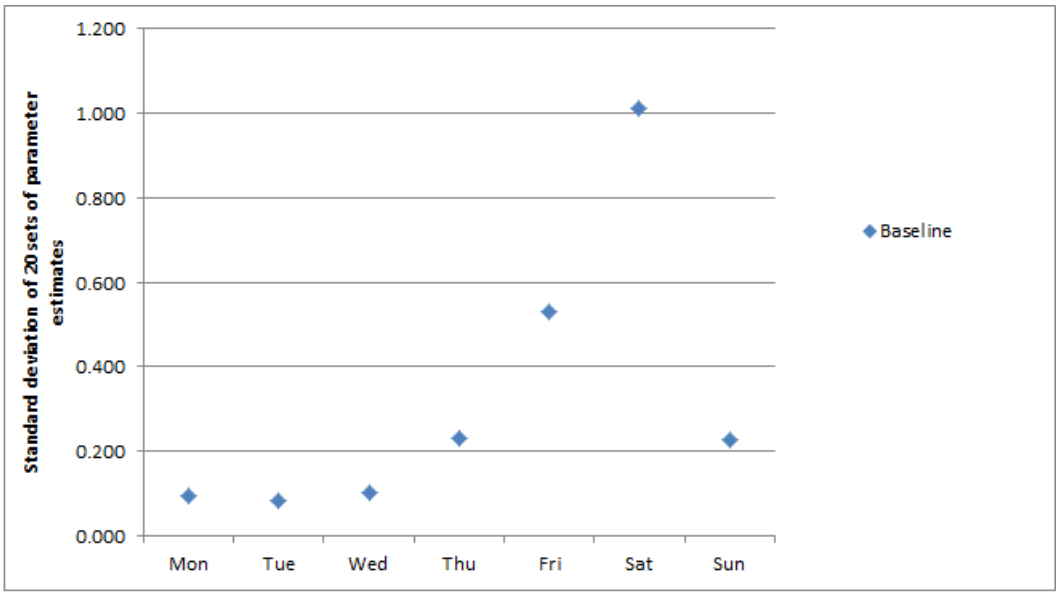


Figure 6.1: Plot showing the standard deviations for the baseline hospital case only.

For each day of the week figure 6.1 shows the standard deviations of 20 runs of the calibration process. The standard deviations from Monday to Wednesday are relatively small, approximately 0.1, and the average parameter estimates for

these days are 2.08, 2.41, and 2.56 respectively. This indicates that the parameter estimates are reasonably consistent over the 20 runs with the calibration process being quite accurate for these days.

Friday's and Saturday's standard deviations on the other hand are the worst of the week, greater than 0.5 and 1.0 respectively. This is at odds with the other results of the week, with Thursday and Sunday's standard deviations being slightly greater than Monday to Wednesday but still relatively small. This may be a consequence of the small volume of arrivals on Friday and Saturday compared to the rest of the week.

We anticipated that small volumes of arrivals would lead to poorer accuracy from the calibration process which is consistent with Friday's and Saturday's results. However this does not explain why Sunday's standard deviation is smaller than Friday's despite having fewer arrivals. Clearly arrivals alone does not dictate the accuracy of the calibration process but other considerations such as length of stay parameter values and possible interactions between days need to be accounted for. We therefore explore this further in the following sections of this chapter.

6.3.2 Changing the volume of arrivals

Figure 6.2 shows the standard deviations for each day of the week for five different scenarios. As the volume of arrivals increases from the "halved arrivals" case to the "baseline" and the "doubled arrivals" we observe the standard deviation of the parameter estimates decreases in the majority of cases for all days of the week. This is as we expected that the larger the volume of arrivals the better the calibration process performs by producing more consistent parameter estimates.

We note Saturday and Friday have the greatest standard deviations for the "baseline", "halved", and "doubled" cases due to having so few arrivals (Saturdays have the fewest arrivals of the week and Fridays the fewest of the weekdays). But when we scale up the weekend arrivals the increasing pattern of standard deviations continues to Friday (where the arrivals are unchanged) while Saturday's standard

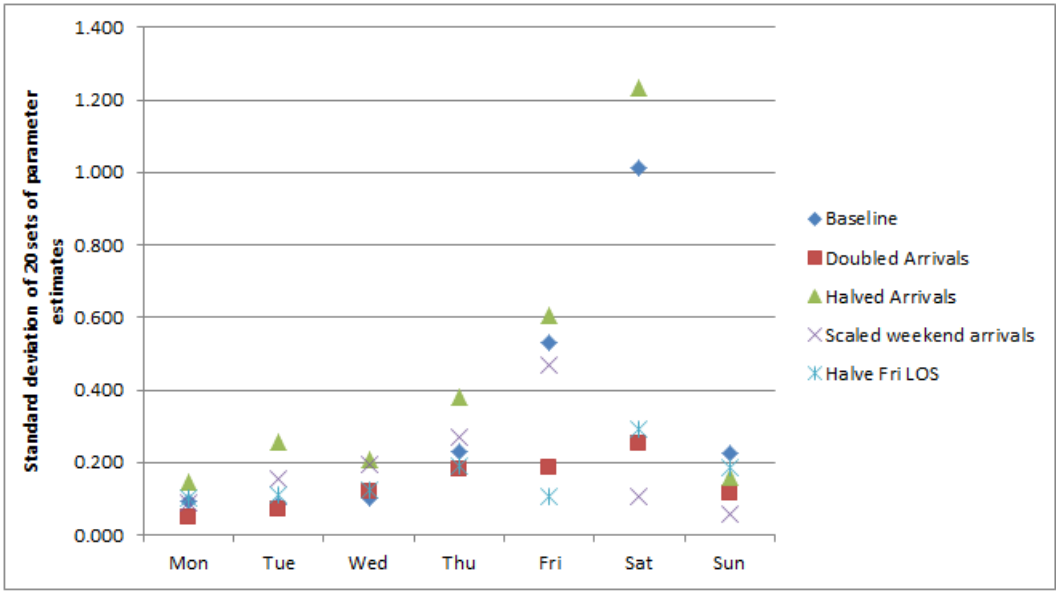


Figure 6.2: Measuring the impact of arrival volume on the calibration process.

deviation is much smaller (Figure 6.2). This all indicates increasing the volume of arrivals improves the accuracy of the calibration process by reducing the standard deviation of the parameter estimates.

An important case to highlight is Saturday's standard deviation when we halve Friday's length of stay (Figure 6.2 the "Halve Fri LOS" case is the same as the "baseline" in all respects except the true Friday length of stay parameter has been halved). Note that in this scenario Saturday is unchanged from the "baseline" case. But it now has a standard deviation comparable to that obtained when doubling the arrivals. Thus changing Friday's length of stay parameter value has affected not only Friday's standard deviation but Saturday's as well. This indicates that the accuracy of the calibration process is not independent between days, but that the accuracy on each day can affect that on others.

6.3.3 Changing length of stay parameter size

When the length of stay parameters are doubled the standard deviation for each day increases ("double LOS" case in Figure 6.3). This pattern remains consistent when we double only some of the length of stay parameters, the standard deviations

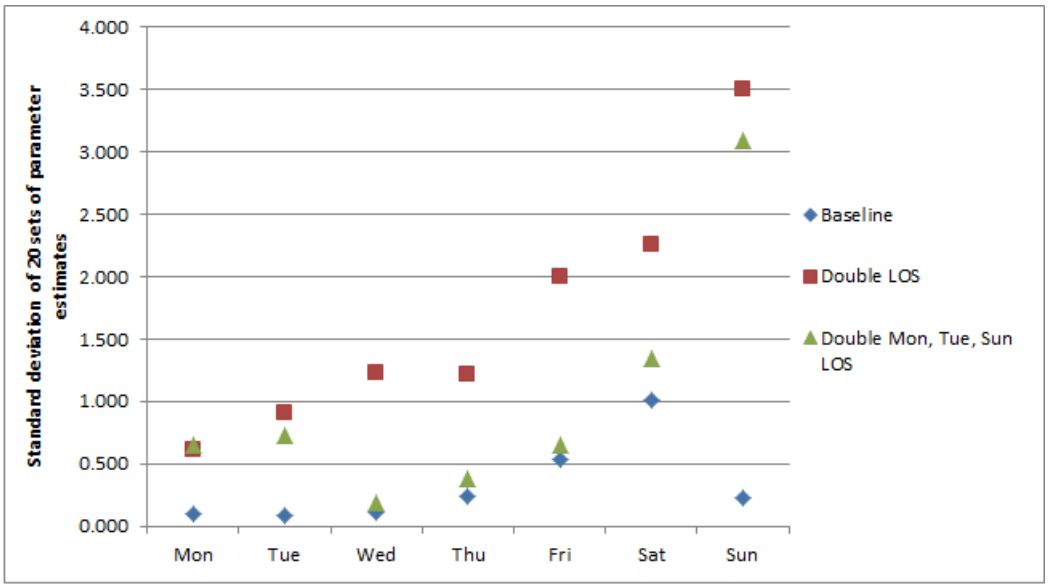


Figure 6.3: Measuring the impact of length of stay parameter size on the calibration process.

on those days are similar to the "double LOS" case while the standard deviations on the days which were unchanged are approximately the same as the "baseline" case. This is consistent with larger length of stay parameter values being more difficult to estimate in the calibration process.

We also note a large difference between Sunday's standard deviations in the "baseline" and "double LOS" cases, whereas in contrast Monday has a relatively small difference between its results. This is a consequence of the relative volume of arrivals affecting the impact of the length of stay size. The small change in Monday's standard deviations is a result of the small volume of arrivals on Saturday and Sunday. When the length of stay parameter is doubled the small volume of arrivals at the weekend results in a relatively small impact on Tuesday's and Wednesday's occupancy (which affects Monday's parameter estimation) - therefore, for Monday's parameter estimation, the difference between the "baseline" and "double LOS" cases is relatively small. But Sunday's standard deviation suffers significantly due to the relatively large volume of arrivals on Thursday and Friday.

6.3.4 Changing size of data

The calibration process works by comparing the modelled daily occupancy and the simulated daily occupancy. So changing the size of the data in this context means, with the three scenarios defined in Table 6.3, and the results shown in Figure 6.4.

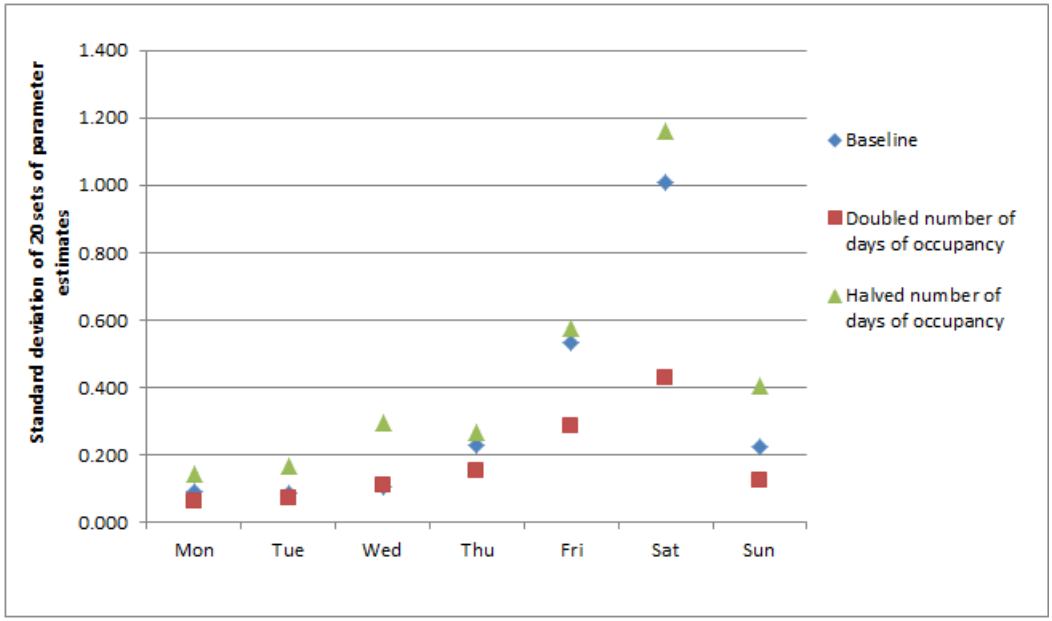


Figure 6.4: Measuring the impact of data size on the calibration process.

The standard deviations are smaller for all days of the week when using more data but the degree of the impact when halving and doubling the data varies by day of week. The improvement in the standard deviation when doubling the data size appears minimal from Monday to Thursday where the variability is at its lowest. Friday and Saturday exhibit the largest standard deviations of the week and doubling the data has the greatest impact on those days in terms of reducing the standard deviation. Halving the data has increased the standard deviations but has had a relatively subdued impact on Friday's, Saturday's and Sunday's standard deviations compared to doubling the data. This indicates there are limits to the impact of the amount of data can have on our estimation of the length of stay parameters.

6.4 Discussion

This chapter introduces a method of calibrating the analytical model, estimating day-of-week length of stay parameters given arrival and occupancy data. However this approach is subject to sampling variation and we therefore investigate how the volume of arrivals, the length of stay parameter values and the data size all impact on the accuracy of the calibration process.

Our results show that larger volumes of arrivals and smaller length of stay parameter values can generally reduce the variability of the parameter estimates. Another factor we consider is data size, where increasing the amount of data improved the calibration process. We observe that halving or doubling the data has a minimal impact on certain days while the greatest improvement in accuracy when doubling the size of the data came on Saturday, where the variability was greatest to begin with. Typically though for the majority of days the standard deviations remains relatively small even when halving the amount of data.

We have also found evidence that other factors affect the accuracy of the calibration process. The distribution of the arrivals over the week and differing length of stay parameter values between days both impact on the calibration process. For example we earlier stated the larger the length of stay parameter values, the better the accuracy of the calibration process however we must also consider the size of the arrivals on each day as this affects the size of the impact of changing length of stay parameter values. The accuracy of the calibration process is not independent between days.

The purpose of investigating the calibration process is to provide some confidence that the parameter estimates used in Chapter 5 are reasonable and do not dramatically affect the results drawn from the model for any given hospital. Clearly the modelled occupancy results (and therefore the performance indicators) are subject to some sampling variation. However since the performance indicators produced are intended as a tool for indicating the performance of hospitals and encouraging further investigation, the process of further investigation should insure

against any over-reaction to possible sampling errors.

Chapter 7

Conclusions

This thesis aims to develop the use of infinite server models for characterising patient flows in healthcare systems in order to improve and inform decision making. This is done in two ways: firstly we extend the existing theory of infinite server models in particular relaxing assumptions on the arrivals distribution when deriving formulae for the mean and variance of demand, and secondly we offer alternative approaches for exploiting the model outputs to improve hospitals. A consequence of the second point, when using the infinite server model to characterise elective patient flow (Chapter 5), was the need for a method to estimate the parameters of the day-of-week dependent length of stay distributions. This method is subject to sampling variation for which we provide a statistical analysis in Chapter 6.

7.1 Summary of thesis findings

In Chapter 2 we summarise and present important results for both single node infinite server models and networks of infinite server queues. This includes closed-form formulae for the mean and variance of demand which can be derived using these models. We also present some interesting examples of infinite server models used in a healthcare context.

In particular we draw attention to three papers (Gallivan and Utley (2005),

Helm and Van Oyen (2014), Bekker and Koeleman (2011)) which choose to use optimisation alongside their infinite server models to determine optimal elective admission schedules. We highlight the limitations in the current applications of infinite server models and suggest an alternative use of the model outputs in the form of performance indicators.

Chapter 3 builds on the theoretical results defined in Chapter 2 and extends the theory of single node infinite server models. We are able to relax assumptions placed on the arrivals of the single node infinite server model to incorporate a general arrival distribution. This less restrictive assumption on arrivals makes the infinite server model more flexible and able to incorporate different types of arrivals while retaining closed-form expressions for the mean and variance of demand.

In this chapter we also note Massey and Whitt (1993)'s relatively difficult to understand methodology for computing the mean demand of individual nodes in a network system of interconnected queues. We present a simpler and more intuitive approach for computing the mean demand of individual nodes in the network system and show this result is equivalent to Massey and Whitt (1993)'s.

Chapter 4 introduces the emergency inpatient work of 30 hospitals and demonstrates the use of the infinite server model for characterising emergency patient flows. We develop performance indicators based on the modelled results to provide hospital size adjusted statistics for comparing the performance of the hospitals. We suggest ways in which the model-based performance indicators can be used to help decision makers improve the management of hospitals and discuss which hospitals are potentially performing poorly or well.

Chapter 5 introduces the counterpart elective inpatient work of the same 30 hospitals as in Chapter 4. This chapter demonstrates the different modelling challenges required for elective work, including focusing on weekday only patient demand (excluding weekend demand), using a general arrival distribution, and day-of-week dependent length of stay distributions.

We are able to construct model-based performance indicators for the elective

workload and the total workload (the combined elective and emergency work). This chapter highlights the importance of including both elective and emergency workloads when analysing hospital performance, though it should be noted the model assumes elective and emergency patients are treated independently which may not necessarily be realistic. We use the model-based performance indicators for emergency, elective and total work to analyse the hospitals and help inform decision makers by potentially identifying good and poor management practices

Chapter 6 addresses a statistical issue which arises from incorporating day-of-week dependent length of stay distributions for elective work in Chapter 5. We estimate the parameters of the day-of-week length of stay distributions using a calibration process which is subject to sampling variation. In this chapter we develop and apply a method to investigate the factors which affect the parameter estimation, namely the volume of arrivals, the length of stay parameter values, and the size of the data and provide a statistical analysis of the accuracy of the calibration process subject to these factors. The purpose of this analysis is to provide some confidence that the parameter estimates used in Chapter 5 are reasonable and do not significantly affect the conclusions drawn from the modelled results.

This thesis has contributed to the aim of helping improve and inform decision making in healthcare systems by using infinite server models to characterise patient flows. We have detailed how infinite server models are useful in describing healthcare systems and have extended the present theory of these models. Current healthcare applications of infinite server models are limited in their use of the model outputs, and we offer a different approach by constructing model-based performance indicators to compare hospitals based on their performance.

7.2 Future directions

Possible directions for further work include development of further queueing theory, development of further model-based performance indicators, and the application of model-based performance developed in this thesis.

In Chapter 3 we offer an alternative method for computing the mean demand of individual nodes in a network of infinite server queues, however it remains to be shown whether a similar result can be found for the variance of demand. Massey and Whitt (1993) do not present results for the variance in the general arrival case and nor do Helm and Van Oyen (2014) though they manage to derive expressions for the variance for individual wards and the entire hospital only for the deterministic arrivals case. Variability of demand is an important statistic for determining hospital performance and could be valuable when constructing performance indicators for a network healthcare system.

It would be of interest to apply the theory in Chapter 3 to construct performance indicators for networks of nodes in a similar fashion to that done in Chapters 4 and 5 for the single node case. The network theory may prove useful in modelling multiple wards of a hospital or in individual departments for example an A&E department where patient pathways take them through different workstations.

Accounting for the network nature of hospitals is valuable because it acknowledges the various patient routes and potential areas for bottlenecks. Applying the infinite server theory to network systems may provide different challenges to the single node case. For example multi-dimensional performance indicators may be required to capture the conflicting priorities of different wards in a hospital.

Chapters 4 and 5 demonstrate model-based performance indicators being used to rank hospital performance in order to help inform decision makers of which cases require closer inspection. Such investigations at individual hospitals might then uncover poor practices and hence lead to improved performance.

However in research terms a more valuable contribution might be to use the model-based performance indicators to identify all the best (or worst) performing hospitals together. A co-ordinated approach analysing the best (or worst) performing hospitals could then reveal common traits and underlying properties among the hospitals.

This work could also act as a test for how effective or informative the model-based performance indicators are. For example it may be the case among the hospitals which ranked poorly according to the performance indicators that there is a common factor which is adversely affecting the results. In such a scenario it may encourage further refining of the model-based performance indicators to provide more informative results.

In modelling day-of-week dependent length of stay distributions (Chapter 5) we estimate their parameters using a regression based approach which we investigate in Chapter 6. However this demonstrates that the model inputs are subject to sampling error which in turn will affect the accuracy of the performance indicators. This effect is referred to as "input uncertainty" (Henderson, 2003), and there is clearly scope for further investigation of the input uncertainty associated with model-based performance indicators.

Appendices

Appendix A

Comparison of the variance of bed demand

Building on the idea of the Gallivan and Utley model being a special case of Helm and Van Oyen's model we show the variance expressions for both the elective and emergency patient demands are equal. The main differences between the expressions for the variance are the use of different notation and the expansion to multiple wards for Helm and Van Oyen's model.

A.1 Variance of elective bed demand

Since both Helm and Van Oyen (2014) and Gallivan and Utley (2005) consider elective and emergency patients separately we begin by focusing on the elective case only. Table A.1 summarises the different notation used for the variance of elective bed demand in each paper.

We begin with Helm and Van Oyen's definition of the variance of the elective bed demand in ward u, $\sigma^2_{d_1,u}(\Theta)$, where $d(n) = d_1 - d_2 + 7(t-n)$ and show this is

Helm and Van Oyen	Gallivan and Utley
d_1 = day of the planning cycle	d
d_2 = day of the cycle a patient is admitted	i
u = ward	-
k = patient type	h
W = set of hospital wards	-
D = set of patient types	H
Θ = elective admission schedule	N
N = planning horizon (for a planning horizon of a week N equals 7)	C
$p_{s,k,u}(t-s)$ = the probability a patient of type k who arrives at time s is in ward u at time t	$p^h_{s,t-s}$

Table A.1: Table summarising the different notation used in Helm and Van Oyen (2014) and Gallivan and Utley (2005) for the variance of elective patient demand.

equal to the equivalent expression derived in Gallivan and Utley (2005):

$$\sigma^2_{d_1,u}(\Theta) = \sum_{d_2=1}^{7}\sum_{k\in D}\Theta_{k,d_2}\sum_{n=0}^{\infty}p_{d_2+7n,k,u}(d(n))(1-p_{d_2+7n,k,u}(d(n))) \tag{A.1}$$

$$\sigma^2_{d_1}(N) = \sum_{d_2=1}^{7}\sum_{h=1}^{H}N_{h,d_2}\sum_{n=0}^{\infty}p^h_{d_2+7n,d(n)}(1-p^h_{d_2+7n,d(n)}) \tag{A.2}$$

$$\sigma^2_{d}(N) = \lim_{t\to\infty}\sum_{h=1}^{H}\sum_{i=1}^{7}N_{h,i}\sum_{n=0}^{t}p^h_{(i+7n),(7(t-n)+d-i)}(1-p^h_{(i+7n),(7(t-n)+d-i)}) \tag{A.3}$$

$$= \lim_{t\to\infty}\sum_{h=1}^{H}\sum_{i=1}^{7}N_{h,i}\sum_{w=0}^{t}p^h_{i,(7w+d-i)}(1-p^h_{i,(7w+d-i)}) \tag{A.4}$$

$$= \sum_{h=1}^{H}\sum_{i=1}^{7}N_{h,i}\sum_{w=0}^{\infty}p^h_{i,(7w+d-i)}(1-p^h_{i,(7w+d-i)}) \tag{A.5}$$

where Equation (A.5) is the variance of the elective patient demand in Gallivan and Utley (2005). Since we can think of Gallivan and Utley's model as the single ward case of Helm and Van Oyen's model, the first step drops the index u which represents the various wards.

Equations (A.3) and (A.4) are equal because the indices of the length of stay probabilities are equal. If we take the first index of the length of stay probabilities then using the notation in Gallivan and Utley, i, indicates the day of the cycle a patient arrived but not which cycle. On the other hand in Helm and Van Oyen, $i+7n$, indicates not only the day of arrival but also which cycle the patient arrived.

In this case we assume the length of stay distribution does not depend on which week a patient arrived so setting the first index to i is no different to $i + 7n$. The second index differs only in the manner in which the number of weeks that a patient may be resident for are counted. They are represented as $t-n$ by Helm and Van Oyen, and w by Gallivan and Utley. In Helm and Van Oyen's model t acts as the limit to the number of weeks a patient may remain in the system i.e. when n exceeds t the second index is negative, for which the length of stay distribution is 0 by definition. Similarly in the Gallivan and Utley case if a limit to the number of weeks a patient can remain in the system is imposed (using the same notation in Helm and Van Oyen (2014), this would be t) then when the number of weeks a patient remains in the system exceeds this limit i.e. when w exceeds t, then by definition the length of stay distribution is always 0.

A.2 Variance of emergency bed demand

Showing the expressions for the variance of the emergency bed demand in Helm and Van Oyen's and Gallivan and Utley's models are equal follows a similar pattern to the elective case.

Helm and Van Oyen	Gallivan and Utley
τ_k = maximum length of stay for a patient of type k	-
j = ward	-
k = patient type	$h = 0$
t = time of the planning cycle	d
s = time of arrival	i
$p_{s,k,u}(t-s)$ = the probability a patient of type k who arrives at time s is in ward u at time t	$p^h_{s,t-s}$
$\alpha_k(s)$ = arrival rate function for a patient of type k who arrives at time s	$N_{k,i}$

Table A.2: Table summarising the different notation used in Helm and Van Oyen (2014) and Gallivan and Utley (2005) for the variance of emergency patient demand.

Helm and Van Oyen derive the mean of the emergency patient demand in each ward. Since the emergency patient demand (i.e. the number of emergency

patients) in each ward follows a Poisson distribution, the variance of the emergency bed demand is equal to the mean. Therefore the variance of the emergency patient demand in ward j is:

$$m_j(t) = \sum_{k=1}^{n} \int_{t-\tau_k}^{t} \alpha_k(s) p_{s,k,j}(t-s) ds \tag{A.6}$$

$$m(t) = \int_{t-\tau}^{t} \alpha_0(s) p^0_{s,t-s} ds \tag{A.7}$$

The first step uses the same idea when showing the variances are equal in the elective case. Again the fact that Gallivan and Utley's model is the single ward case of Helm and Van Oyen's allows us to drop the j index representing the wards. In addition since Gallivan and Utley only consider one emergency patient type, which they indicate with $h = 0$, the k index, which represents the multiple emergency patient types in Helm and Van Oyen's model, is set to 0.

Since Gallivan and Utley assume the rates of arrival follow a weekly pattern the arrival rate functions from $t - \tau$ to t can be written in terms of the Poisson means, $N_{0,i}$. For $x \in \mathbb{N}$:

$$\begin{aligned} \alpha_0(1) &= \alpha_0(1+7x) = N_{0,1} \\ &\vdots \\ \alpha_0(7) &= \alpha_0(7+7x) = N_{0,7}, \end{aligned}$$

where in Gallivan and Utley the number of emergency admissions on day i of the planning cycle is assumed to be Poisson distributed with mean and variance, $N_{0,i}$.

Helm and Van Oyen write the variance function as an integral whereas Gallivan and Utley write the variance as a summand. This choice reflects Helm and Van Oyen's use of an arrival rate function in continuous time, in this sense their formulation can consider any choice of arrival rate function. Gallivan and Utley's form is simply a special case where their "arrival rate function" is a step function, whereby the arrival rate $N_{0,i}$ changes at discrete time points (i.e. each day).

A second point to consider centres around the length of stay distributions. In a similar vein to the arrival rate function, Helm and Van Oyen's length of stay distribution is a general function which is continuous over time. Again Gallivan and Utley's distribution is a special case of Helm and Van Oyen's, akin to the arrival rate function, where their length of stay distribution $p_{i,7w+d-i}$ is a step function changing at discrete time points.

Therefore we can change from the integral in Helm and Van Oyen's model to a summand in Gallivan and Utley's.

$$m(t) = \sum_{s=t-\tau}^{t} \alpha_0(s) p^0_{s,t-s} \tag{A.8}$$

$$= \sum_{i=1}^{7} \sum_{w=0}^{\infty} N_{0,i} p^0_{i,7w+d-i}, \tag{A.9}$$

where Equation (A.9) is the contribution to the variance of bed demand from emergency patients in Gallivan and Utley (2005).

Since we have already shown how the arrival rate function, $\alpha_0(s)$ can be written in terms of the Poisson means, $N_{0,i}$ above, the only apparent difference between equations (A.8) and (A.9) lies in the indices of the length of stay probabilities. In Helm and Van Oyen's model the first index indicates the day of arrival and which cycle a patient arrives. While in Gallivan and Utley the first index does not indicate which cycle a patient arrived. Additionally in Helm and Van Oyen's notation the first index does not necessarily begin at 1 but at $t - \tau$, meaning patients only contribute to the variance of bed demand if they arrived on day $t-\tau$ or later. Gallivan and Utley achieve the same result by setting the persistence distribution for patients who arrived earlier than day $t - \tau$ to 0. For instance if the day of interest is $t = 16$ and patients may be resident for up to $\tau = 13$ days, patients who arrived before day $t - \tau = 3$ do not contribute to the variance. So for Gallivan and Utley, in this case, $p^0_{1,7w+d-i} = p^0_{2,7w+d-i} = 0$.

The duration a patient may stay in Helm and Van Oyen's model ranges from 0 to τ. The second index equals 0 in Gallivan and Utley (2005) when $w = 0$ and

$i = d$. And for values of w where the second index exceeds τ i.e. the duration exceeds the length of time a patient may remain in the system, the persistence distribution is 0 by definition.

We have shown separately for both the elective and emergency contribution to the variance of bed demand that the variance in Gallivan and Utley (2005) is a special case of the Helm and Van Oyen (2014) equations for the variance of bed demand.

Appendix B

Launceston General Hospital

Progress Report

Data Analysis and Preliminary Modelling for Launceston General Hospital

This report describes the aims of our preliminary models, the assumptions and transformations made to the raw data in order to derive a usable data format, the results of analysing this transformed data, and the structure and results of the basic mathematical/computational models used to estimate bed demand at Launceston General Hospital. We conclude by running alternative scenarios ("What-Ifs") through the models.

David Oakley and Daniel Suen (Supervisory Team: Dave Worthington, Stephan Onggo, Matt Sperrin)

10/13/2012

10/13/2012

OVERVIEW

Phase 1 of the Patient Flow Modelling project for Launceston General Hospital is to develop and parameterise an uncapacitated model of patient flows at LGH. The aim of such uncapacitated models is to estimate the 'ideal' resource requirements of different patient flow scenarios, taking into account user-defined activity levels; and experimenting with them will lead to a better understanding of the problems and potential solutions facing LGH management.

The anticipated benefits of the uncapacitated modelling phase are:

- The requirements for detailed information on hospital systems and for detailed data are relatively modest;
- It nevertheless provides estimates of desirable resource levels under different scenarios, which can be compared with actual (or planned) resource levels to identify likely pressure points and bottlenecks;
- It provides guidance on resource levels needed to improve the balance of existing or planned resources;
- It will provide a natural baseline model for later phases of modelling, designed to look in more detail at the impact of fixed resource levels at key points in the system;
- By providing an overview of patient flows and likely bottlenecks, it will facilitate the selection of areas of activity where further modelling is worthwhile.

Because uncapacitated models cannot fully represent the consequences of the congestion caused by limited resources, it is envisaged that phase 2 of the project will be to model those parts of the system where this is particularly important in greater detail. Hence whilst phase 1 is primarily concerned with getting maximum value out of the uncapacitated models, it is also important that phase 1 recognises and highlights those issues where the capacitated models of phase 2 are needed.

Data arrived from LGH early in July, together with a data dictionary, and this report summarises early analyses of this data set and some early modelling work using parameters derived from the data. The purpose of this progress report is to highlight issues arising to date on which guidance is required. These issues can be grouped under the following headings:

i. Are we interpreting the data correctly?
ii. What levels of granularity and focus are likely to be useful to hospital management at LGH? Our initial analysis has removed day cases and renal patients, and then divided the remaining patients into electives and non-electives. Is there value in dividing patients into smaller subsets? Should day cases be including for some issues? Etc.
iii. What sorts of decisions and issues would LGH like the project to inform?

The analyses and modelling reported are deliberately provisional and naive, and all the results reported at this stage are for validation and demonstration purposes only. {We believe that it will be much more productive to go forward in this way rather than to try to ask many individual questions at this early stage. One of the strengths of our approach in

phase 1 is that it should be relatively easy to rerun analyses and models in the light of corrections and clarifications that emerge as a result of this first progress report.}

The areas addressed in the attached appendices are:

- Data exclusions and assumptions (see section 1);
- Analysis of historical arrival and length of stay patterns (see section 2);
- Demonstration of model using historical data (see section 3);
- Demonstration of 'what if .....' use of model (see section 4).

In response to this first progress report we ask that LGH:

i. Advise us of corrections we need to make to our understanding and analysis of the data.
ii. Comment on our current inclusions and exclusions of patients, and on the level of granularity at which patients should be included.
iii. Comment on the extent to which our analysis of the historic data seems valid? Would it be more meaningful to focus on different patient groups to those chosen so far? Are the time segments that we have chosen sensible?
iv. Suggest specific issues going forwards where the 'what if ..' modelling approach could be applicable.
v. Comment on any other issues they think important.

Finally we note that the theatre requirements of patients have not yet been incorporated into our analyses and modelling. This aspect of phase 1 is ongoing.

10/13/2012

PROGRESS REPORT

Data Analysis and Preliminary Modelling for Launceston General Hospital

Prior to receiving data from LGH, the authors have been involved with developing a mathematical and a computational model of some hypothetical hospital, with the hope that these models could later be used to describe bed demand for a real hospital, given suitable data. In theory, estimates of bed demand derived from either modelling approach are equivalent; however these methods have individual strengths and weaknesses.

Until recently, the parameters used in the hypothetical models to describe say, patient arrival patterns, had been based on rough estimates provided to us from a telephone conference held in February. Estimates of these parameters can now be refined, based on the datasets provided, and the analysis carried out by the authors.

The structure of this report is as follows. Firstly, we describe the data we deemed useful and how it was used. This is to ensure that we have not used data inappropriately and also so any assumptions made during this process are clearly communicated. Secondly, the data is analysed to determine the presence of any trends or patterns. This process has the ancillary use of providing the parameters for the two models. Thirdly, we show reasonable agreement between our estimates of bed demand and actual bed demand derived from the datasets provided. Finally, alternative scenarios (so-called “What-Ifs”) are run through the two models to show the impact on the estimates of bed demand and to demonstrate how the models can be used.

1. DATA PROCESSING

The data received spans a period beginning at 01/01/2010 until 30/06/2012, recording patients who were admitted and discharged in this time. In addition patients admitted before 01/01/2010 but were still residing in the hospital on 01/01/2010 were included.

An initial study of the data revealed some key categories likely to be important in the analysis of the data and the implementation of our model. In particular the “AdmissionTypeRefId" variable divides the patients into three main categories; “Urgency status assigned - elective", “Urgency status assigned - emergency" and “Urgency status not assigned". There are over 90000 cases in the data of which 55000, 24000, 12000 were labelled elective, emergency and not-assigned types respectively. The remaining five admission types made up less than 3500 patient cases.

In addition, the majority of not-assigned patients appear to be maternity related cases. This is based on the fact that of the not-assigned patients nearly 70% arrived on one of two wards; Ward4B (6000) and Ward4O (2000). Of the patients arriving in those two wards almost all were admitted under the specialty type “obstetrics" or “paediatric medicine". A potential

reason for maternity cases being assigned this status type is because they are not emergency patients, nor do they arrive from a waiting list meaning there may only be limited scope for managing patients of this type. With this in mind the later analysis excludes the not-assigned type and focuses on the two larger patient groups; elective and emergency types.

1.1 Pre-Processing Steps

The following changes have been made when working with the data:

- All cases are divided into one of two categories by "type of admission"; "urgency status -elective" and "urgency status - emergency".
- Renal patients are removed by "admission specialty" (as opposed to removing by say "admission ward"). Part of the reasoning for this is because they contribute such a large number of patients (26000 or 28% of all cases), we can possibly assume some sort of independent setup for handling these cases.
- Day cases (which make up 52000 or 55% of cases) are removed using the "ActualSameDayFlag" variable. This is in part because we are interested in longer term bed demand.
- 158 patients who arrived before 01/01/2010 were removed to simplify analysis.
- Remove patients with the variable "TotalLOS" equal to 0. With only a small number of cases (103), patients with "TotalLOS = 0" may indicate those that have not yet left the system by 30/06/2012.

The last two bullet points impact on the analysis of patients arriving or leaving at either end of the observed period. This "edge effect" requires special treatment which at this stage has been omitted in favour of simplicity.

1.2 Data Filtering and Transformation

The assumptions listed above form the filters which are applied to the data. Based on our requirements for the models at this stage, the remaining variables of interest to us (shown in Figure 1) are "Inpatient EpisodeId," "Reference Description," "Admission DateTime," "Discharge DateTime" and "Total LOS inDays."

Master Elec/Emer

Inpatient EpisodeId	Reference Description	Admission DateTime	Discharge DateTime	Total LOS inDays
208166	Urgency status assigned - emergency	01/01/2010 04:25:00	02/01/2010 16:56:00	1
208152	Urgency status assigned - emergency	01/01/2010 05:07:00	02/01/2010 15:00:00	1
208129	Urgency status assigned - emergency	01/01/2010 05:38:00	02/01/2010 14:15:00	1
208162	Urgency status assigned - emergency	01/01/2010 06:15:00	02/01/2010 14:48:00	1
208146	Urgency status assigned - emergency	01/01/2010 09:33:00	04/01/2010 15:31:00	3
208151	Urgency status assigned - elective	01/01/2010 10:00:00	08/01/2010 07:30:00	7
208144	Urgency status assigned - emergency	01/01/2010 10:07:00	02/01/2010 17:21:00	1

Figure 1: Excerpt from the master dataset on which all analysis is based.

This dataset can then be used to determine the number of patients arriving on each day, of each type (emergency or elective) by summarising counts of "Inpatient EpisodeId" by "Reference Description" and "Admission DateTime." Similarly, length of stay can be analysed by filtering for either elective or emergency type patients and extracting the "Total LOS inDays" data.

10/13/2012

2. Data Analysis

2.1 Bed Census Analysis

Since our initial aim is to be able to model the number of inpatients present in the hospital, it is useful to recreate this process from the data, without the use of any model. This information forms a benchmark by which the accuracy of any further models can be judged. From the data provided, we can determine the number of patients present in the hospital at midnight on any day between 01/01/2010 and 30/06/2012 for both emergency and elective patients. The resulting time series is plotted in Figure 2.

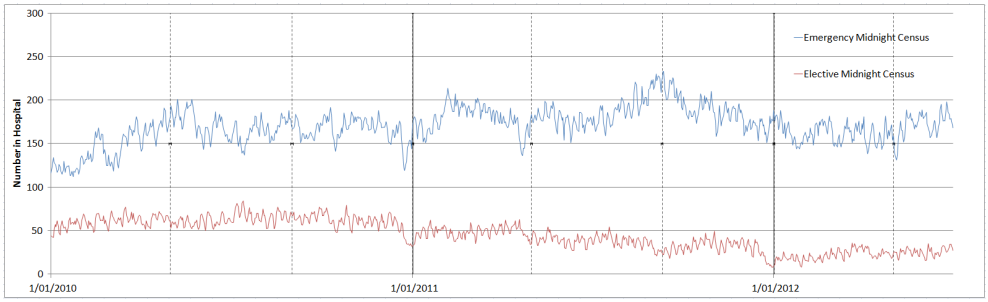


Figure 2: The number of inpatients present at midnight between 01/01/2010 and 30/06/2012.

By considering the elective patient census only, it appears that there is a decreasing trend over time. Plotting the elective census suggested a change in the level every four month period, indicated by the light vertical lines on Figure 2. Therefore when calculating the model inputs, the data was split into seven exclusive time periods of four months, beginning at 1st January 2010 until 1st May 2012.

It is more difficult to assess whether there is a discernible pattern in the emergency series. Further, it doesn't appear to have any upward or downward trend over time. However the census does not appear constant either. In 2011 the average number of inpatients appears to be higher than the average number in 2010, and the recorded portion of 2012. Despite the difficulty in assessing any clear pattern, for now we have divided the data into three periods, indicated by the bolder vertical lines on Figure 2 for each calendar year. Note that, unlike the elective case, the periods are no longer of equal size, due (in part) to the fact that data for the remainder of 2012 is not yet known.

2.2 Patient Arrival Patterns

Focusing on the elective patients first; a difference was detected between the arrival patterns occurring in each of the seven, four month periods. Thus for the elective patients we used different arrival patterns for each of these periods in the two models. Average daily arrivals are shown in Figure 3. Note that for the last period, the average daily arrivals are lower at

least in part caused by removing patients still in the hospital on 30/06/2012 (i.e. an edge effect).

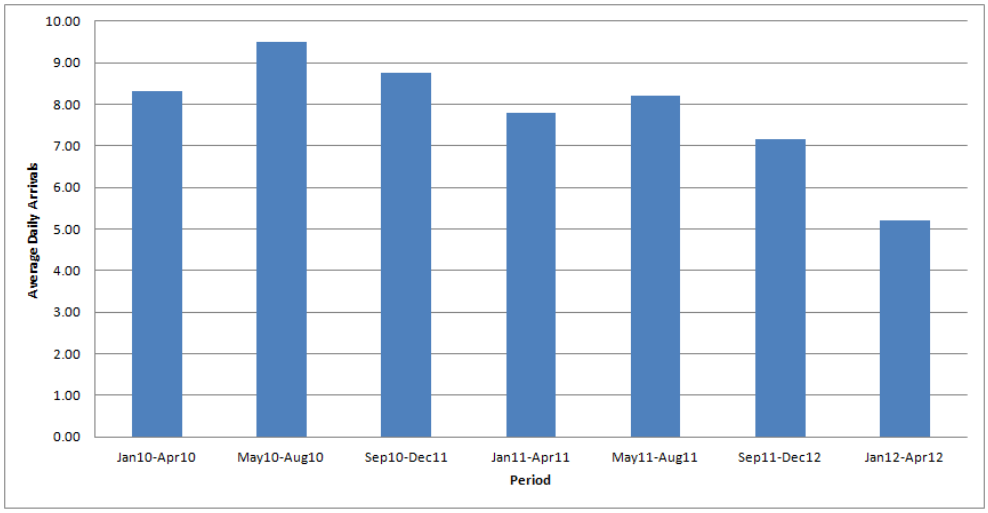


Figure 3: Average daily arrival per period for elective patients.

In conjunction with a changing arrival pattern across each four month period, each day of the week exhibits a different average number of arrivals. As might be expected, there are significantly less elective patient arrivals towards the end of the week. This gives rise to a "day of the week effect" which must be captured together with the "seasonal" effect caused by changes in each of the four month periods. These averages are tabulated and graphed in Table 1 and Figure 4. The results in Table 1 will be used as estimates for the elective arrivals in the models described in Section 3.

	Day						
Period	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Jan10-Apr10	12.18	13.24	12.65	9.53	7.39	1.41	1.82
May10-Aug10	14.17	15.67	14.24	10.29	9.47	1.17	1.83
Sep10-Dec11	12.18	13.65	13.17	9.56	8.89	1.88	1.76
Jan11-Apr11	10.88	13.06	10.18	10.12	8.00	1.06	1.71
May11-Aug11	11.33	13.72	11.11	8.94	9.82	0.82	1.39
Sep11-Dec12	10.06	12.06	11.24	7.67	8.22	0.39	0.88
Jan12-Apr12	6.89	8.88	6.76	6.53	6.18	0.47	0.83

Table 1: Average daily arrivals by day of the week in each period for elective patients.

10/13/2012

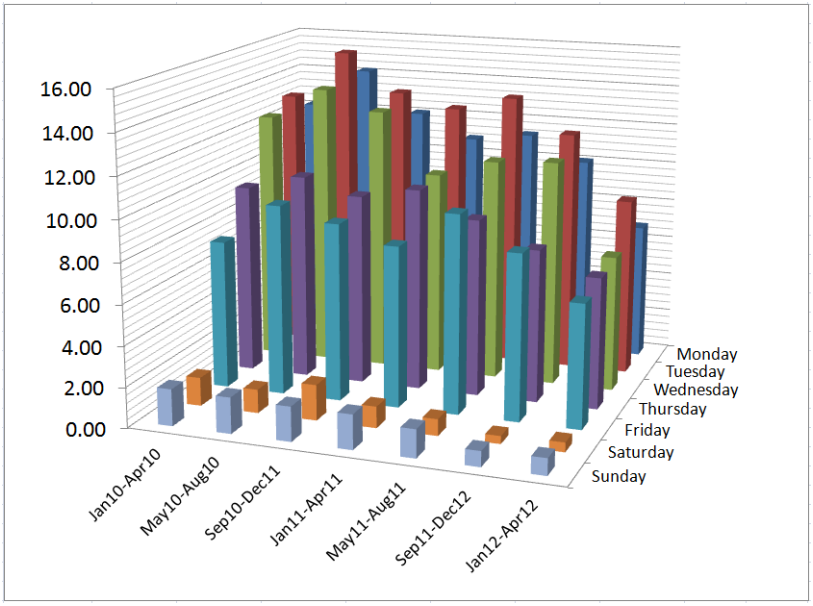


Figure 4: Bar chart of the results in Table 1.

For the emergency patients, a similar process was followed to determine the arrival pattern; however the time span of each period differs. The three periods considered for the emergency patients are 1st January 2010 to 31st December 2010, 1st January 2011 to 31st December 2011 and 1st January 2012 to 30th April 2012.

Figure 5 shows the changing mean daily arrivals for each calendar year for emergency patients. There exists a difference in the average number arrivals occurring during 2010 in comparison to the later years.

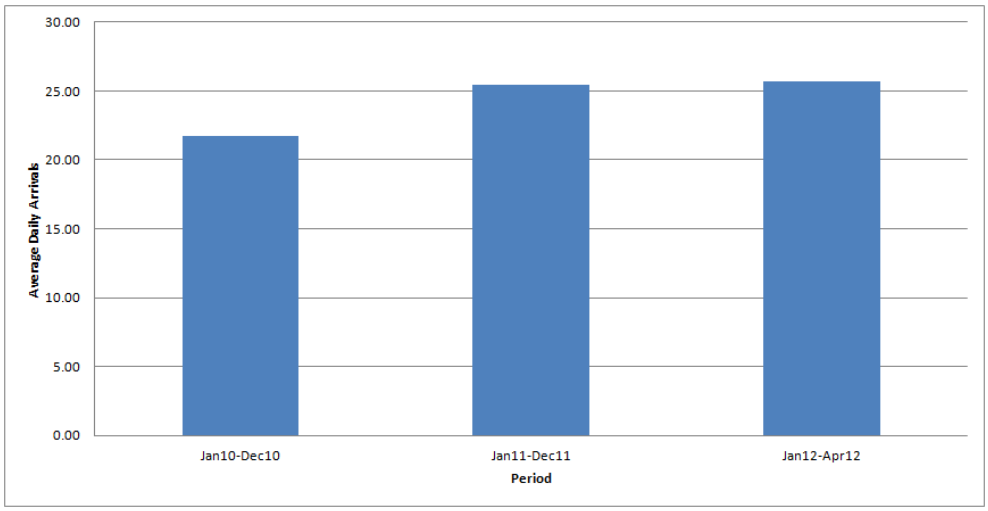


Figure 5: Average daily arrival per period for elective patients.

As for the elective patients, we wish to capture any change in average daily emergency arrivals caused by a day of the week effect along with a longer term year effect. These averages are tabulated and graphed in Table 2 and Figure 6.

	Day						
Period	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Jan10-Dec10	23.79	22.92	23.65	23.33	22.11	17.81	18.21
Jan11-Dec11	27.63	26.65	27.48	26.21	26.88	21.60	21.52
Jan12-Apr12	27.22	26.24	25.71	27.12	28.47	23.41	21.61

Table 2: Average daily arrivals by day of the week in each period for emergency patients.

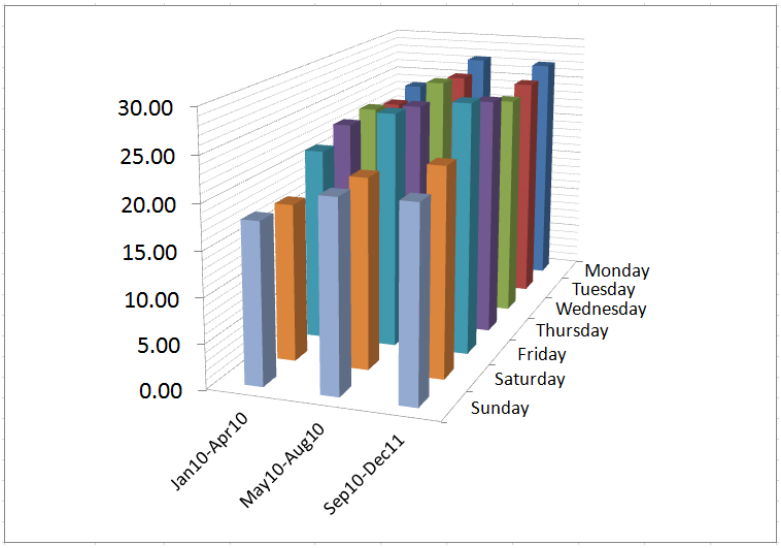


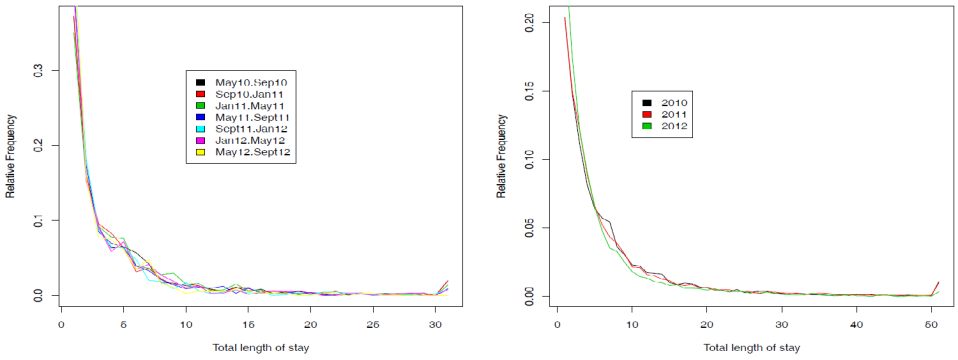
Figure 6: Bar chart of the results in Table 2.

The emergency arrivals show a similar but less pronounced weekly pattern to the elective case, where the number of arrivals is lower at the weekends. The arrivals appear more stable throughout the week rather than peaking at the beginning of the week. Similarly, the results in Table 2 will be used as estimates for the emergency arrivals in our models.

2.3 Lengths of Stay

In contrast to the results for patient arrival patterns, to the naked eye the distributions of patients' lengths of stay shown in Figure 7 seem to be relatively constant over time for the two patient groups (elective and emergency). Closer examination suggests that there may nevertheless be some shifts in the distributions, which are currently being investigated further.

10/13/2012



(a) Elective length of stay densities for each period (b) Emergency length of stay densities for each period

Figure 7: Length of stay plotted against probability of observing a particular length of stay. Elective patients are on the left, while emergency patients are on the right.

In the meantime we have pooled the length of stay distributions across the time periods resulting in the histograms presented as Figures 8 and 9. We used the values from these graphs to determine length of stay in both models.

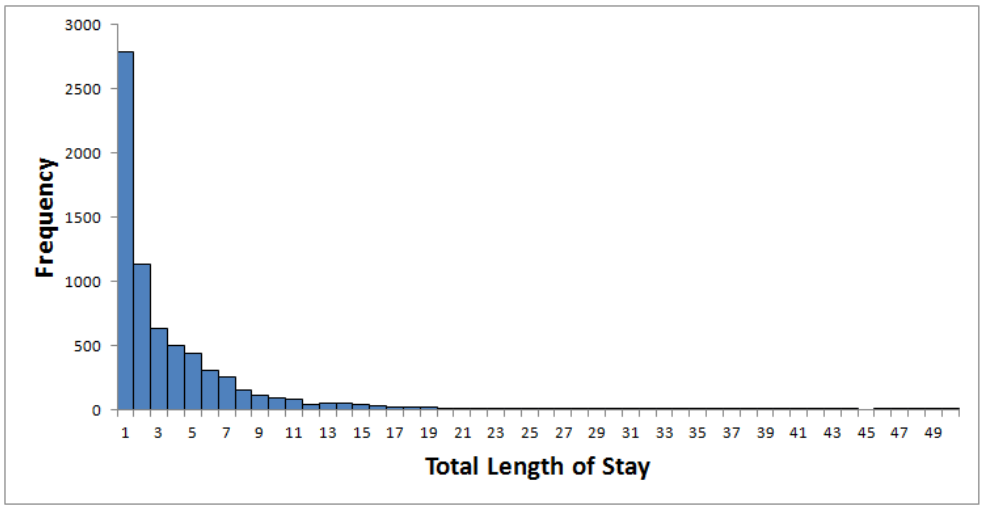


Figure 8: Frequency of elective patient lengths of stay.

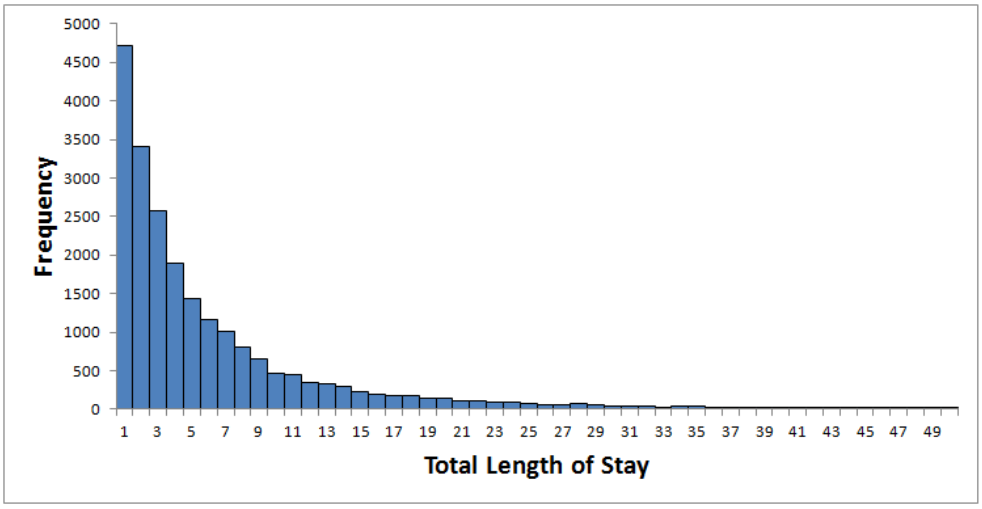


Figure 9: Frequency of emergency patient lengths of stay.

3. Model Validation

3.1 Model Structure

The two main inputs which affect modelling outputs are the rate and pattern of patient arrivals, and the time the patients spend within the hospital itself. This information forms the basis of our model's structure and therefore warrants discussion. As has already been mentioned, the patients have been coarsely grouped into one of two possible categories; elective or emergency. At this stage, no further disaggregation (e.g. into specific specialties) is considered. The reason for this is two-fold. Firstly, we would like to show that the number of patients at any given time can be adequately modelled, if only by using a simple patient grouping. Secondly, the final level of detail has not yet been decided, and this will likely require the feedback and recommendations of the recipients of this document.

Similarly, in our model the hospital itself has not been disaggregated into distinct wards. Again, this is on the basis that we first need to show agreement with the raw data at an aggregate level before further analysis can take place. The resulting conceptual model is represented by the simple process flow diagram in Figure 10. This conceptual model has been applied in both the mathematical model and the computer simulation model.

This diagram is in fact a screen capture from our simulation software. The numbers at each node can be ignored, as these are nothing more than naming conventions employed by the software.

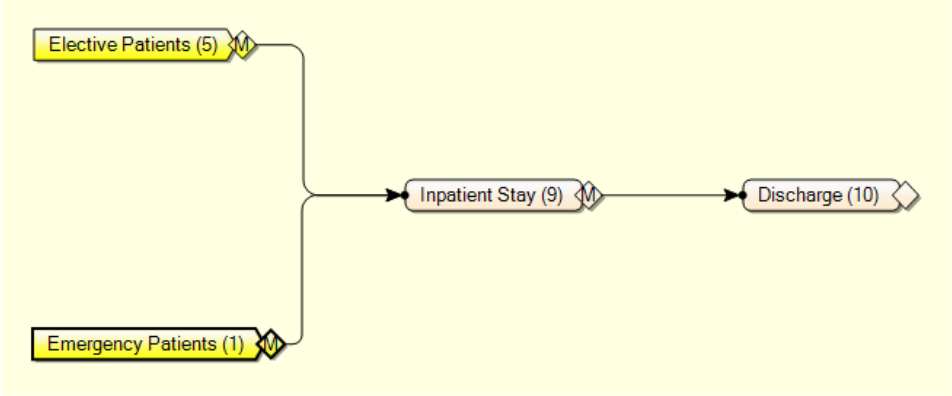


Figure 10: Process flow diagram representing our simple conceptual model.

For our mathematical model (Gallivan and Utley, 2005), the mean and variance of bed demand (this is the mean and variance of the number of beds required in the hospital) are derived using information from the data provided. These functions take into account different lengths of stay for the different patient groups. A weekly cyclic pattern of arrivals for elective patient types is assumed while some standard assumptions are made about the time between successive emergency arrivals.

As has already been mentioned, an alternative to the method proposed by Gallivan and Utley for estimating bed demand statistics is by constructing a computer simulation. Simulations can estimate the mean and variance of bed demand for systems which are more complex than the one proposed in this report; however, they rely on collating data from many simulation runs in order to achieve satisfactory statistical convergence.

3.2 Results

Since this particular model is not complex, and both models produce equivalent results, bed demand estimates from the mathematical model are tabulated in this section. Both sets of results are displayed graphically to illustrate the difference between the estimation methods.

Table 3 shows the mean bed demand results using the parameters based on the data. The mean bed demand is the sum of the contribution from the elective and emergency patient types, as shown in Table 4 and 5. We can see the effect of the reduced weekend elective admissions leading to lower elective patient bed demand at the weekends, while the more consistent emergency arrivals do not share the same pattern. Therefore the lower total bed demand at weekends is caused by the elective rather than the emergency patient type.

	Day						
Period	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Jan10-Apr10	172	179	184	184	181	169	161
May10-Aug10	178	186	191	190	188	174	165
Sep10-Dec11	174	181	186	185	184	171	163
Jan11-Apr11	194	201	203	205	204	191	184
May11-Aug11	195	202	206	205	207	193	184
Sep11-Dec12	190	197	202	201	202	189	180
Jan12-Apr12	184	189	190	192	194	185	178

Table 3: Mean bed demand for all patients; elective and emergency.

	Day						
Period	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Jan10-Apr10	37	42	45	44	42	34	29
May10-Aug10	42	49	52	50	49	39	33
Sep10-Dec11	38	44	47	45	44	36	31
Jan11-Apr11	34	40	41	42	40	32	28
May11-Aug11	36	42	43	42	43	34	29
Sep11-Dec12	31	37	39	38	38	29	25
Jan12-Apr12	22	27	27	27	28	22	18

Table 4: Mean bed demand for elective patients.

	Day						
Period	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Jan10-Dec10	135	137	139	140	140	135	132
Jan11-Dec11	159	161	163	163	164	159	156
Jan12-Apr12	162	163	163	164	167	164	159

Table 5: Mean bed demand for emergency patients.

Figure 11 shows the estimates of the daily elective patient bed demand from the two models alongside the elective census levels, which change in level at each time period of four months (indicated by the vertical lines). The cyclic nature of the estimates from the mathematical model (in green) is because it estimates the underlying (average) levels, whereas the simulation model includes unpredictable daily variations similar to those observed in the census levels. The repeating pattern reflects the repeating weekly nature of arrivals; peaking during the working week while dipping at weekends. A similar pattern, though less marked, can be seen in Figure 12 for the emergency case. Note that the results for the first period, January 2010 - May 2010, and the results after April 2012 have been removed. This is due to the "edge effect" mentioned in Section 1.

It is useful to describe the range of unpredictable daily variations to expect in practice, and this range can also be estimated by the mathematical model. Figure 13 and 14 show the emergency and elective bed census observations varying within a 95% prediction interval (which is cyclic due to the day of the week effect). This can be interpreted by saying that on average, 95 out of 100 of the same type of observations (same day and same period) would be contained within these bounds. Both the elective and emergency bed censes seem to vary within these bounds for most of the observed period.

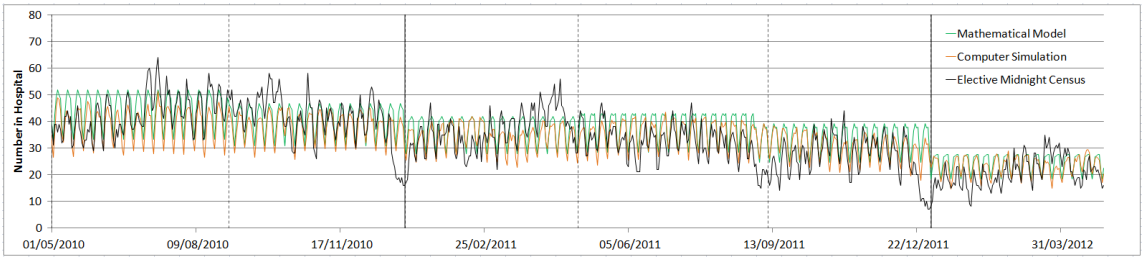


Figure 11: Graph showing the elective census with estimates derived from both models.

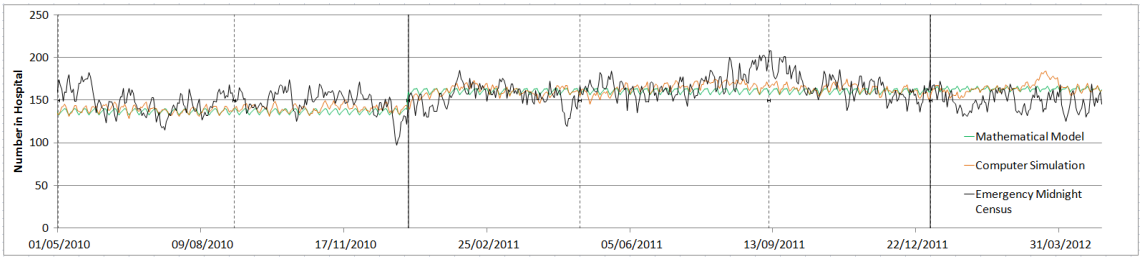


Figure 12: Graph showing the elective census with estimates derived from both models.

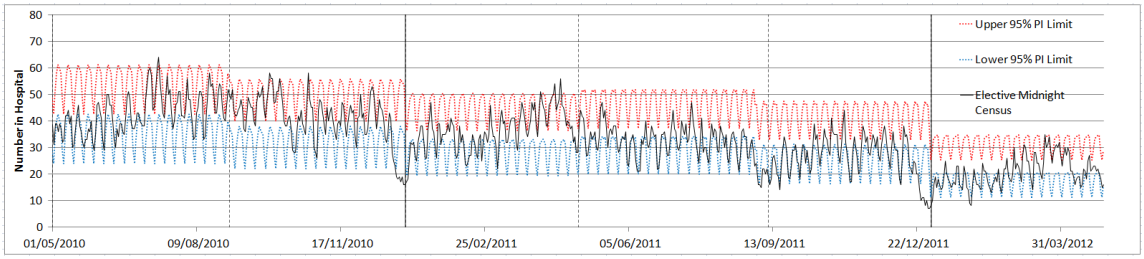


Figure 13: Graph showing the elective patient census and associated 95% prediction intervals.

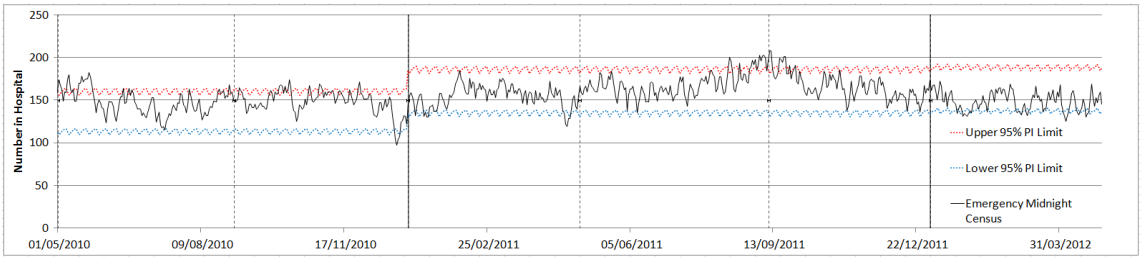


Figure 14: Graph showing the emergency patient census and associated 95% prediction intervals.

10/13/2012

4. Model Experimentation (What-Ifs)

So far, it has been shown that bed demand for different patient types can be adequately modelled. However, this has limited usefulness in itself. The real strength of mathematical/computational modelling is the ability to make changes to the model in order to assess what the impact on the real system might be. Presented below in Figure 15 is the elective bed demand as an output of a computer simulation (shown as the orange series in Figure 11) compared to a simulation of bed demand in which ten elective patients are admitted at the weekend instead of during weekdays. Note that simulation produces the type of random variation seen in real world processes, and therefore the time series differs from the results presented by the mathematical model, in that it is not strictly cyclic. However, analysis of the series can produce estimates of any underlying cyclic process.

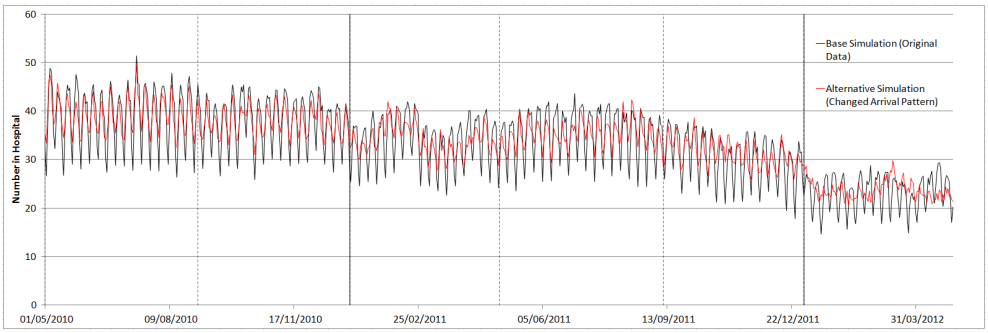


Figure 15: Bed demand estimates from two simulations with differing elective arrival patterns.

Shown in black (Figure 15) is the original simulation in which the model parameters are the estimates taken directly from the data. In red, is the series in which ten elective patients have been admitted during the weekend (five on Saturday, five on Sunday), with two less on each weekday. Even though the weekly workload remains the same, the census shows much less variability under the new admissions policy. Although this may be an unrealistic scenario, it does give some indication of the types of hypotheses which can be tested using simulation models.

5. References

GALLIVAN, S. & UTLEY, M. 2005. Modelling admissions booking of elective in-patients into a treatment centre. *IMA Journal of Management Mathematics,* 16**,** 305-315.

Bibliography

Bard, J. F., Shu, Z., and Leykum, L. (2014). A network-based approach for monthly scheduling of residents in primary care clinics. *Operations Research for Health Care*, 3(4):200–214.

Bekker, R. and Koeleman, P. M. (2011). Scheduling admissions and reducing variability in bed demand. *Health Care*, pages 237–249.

Benes, V. (1957). Fluctuations of telephone traffic. *Bell System Technical Journal, The*, 36(4):965–973.

Bennett, P., Crosbie, J., and Dick, P. (2012). Use of OR by government to inform health policy in england: Examples and reflections. *Operations Research for Health Care*, 1(1):1 – 5.

Bowers, J., Lyons, B., and Mould, G. (2012). Developing a resource allocation model for the scottish patient transport service. *Operations Research for Health Care*, 1(4):84–94.

Brown, M. and Ross, S. M. (1969). Some results for infinite server Poisson queues. *Journal of Applied Probability*, 6(3):604–611.

Bruni, R. and Detti, P. (2014). A flexible discrete optimization approach to the physician scheduling problem. *Operations Research for Health Care*, 3(4):191–199.

Buhaug, H. (2002). Long waiting lists in hospitals: operational research needs

to be used more often and may provide answers. *British Medical Journal*, 324(7332):252–254.

Chow, V. S., Puterman, M. L., Salehirad, N., Huang, W., and Atkins, D. (2011). Reducing Surgical Ward Congestion Through Improved Surgical Scheduling and Uncapacitated Simulation. *Production and Operations Management*, 20(3):418–430.

De Bruin, A. M., Van Rossum, A., Visser, M., and Koole, G. (2007). Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137.

Derlet, R. W., Richards, J. R., and Kravitz, R. L. (2001). Frequent overcrowding in us emergency departments. *Academic Emergency Medicine*, 8(2):151–155.

Eick, S. G., Massey, W. A., and Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Operations Research*, 41(4):731–742.

Farmer, R. and Emami, J. (1990). Models for forecasting hospital bed requirements in the acute sector. *Journal of epidemiology and community health*, 44(4):307–312.

Fomundam, S. and Herrmann, J. W. (2007). A survey of queuing theory applications in healthcare. Technical report, Institute for Systems Research.

Fraser, S. (2010). *Modelling in Healthcare*. American Mathematical Society.

Gallivan, S. and Utley, M. (2005). Modelling admissions booking of elective inpatients into a treatment centre. *IMA Journal of Management Mathematics*, 16(3):305–315.

Gautam, N. (2009). The $M/G/\infty$ Queue. *Wiley Encyclopedia of Operations Research and Management Science*.

Gorunescu, F., McClean, S. I., Millard, P. H., et al. (2002). A queueing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society*, 53(1):19–24.

Green, L. V. (2002). How many hospital beds? *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 39(4):400–412.

Green, L. V. (2008). Using operations research to reduce delays for healthcare. *Tutorials in Operations Research*, 12(5):290–302.

Green, L. V., Kolesar, P. J., and Soares, J. (2001). Improving the sipp approach for staffing service systems that have cyclic demands. *Operations Research*, 49(4):549–564.

Harper, P. R. (2002). A framework for operational modelling of hospital resources. *Health Care Management Science*, 5(3):165–173.

Helm, J. E. and Van Oyen, M. P. (2014). Design and optimization methods for elective hospital admissions. *Operations Research*, 62(6):1265–1282.

Henderson, S. G. (2003). Input modeling: input model uncertainty: why do we care and what should we do about it? In *Proceedings of the 35th conference on Winter simulation: driving innovation*, pages 90–100. Winter Simulation Conference.

Holm, L. B., Lurås, H., and Dahl, F. A. (2013). Improving hospital bed utilisation through simulation and optimisation: with application to a 40% increase in patient volume in a Norwegian General Hospital. *International Journal of Medical Informatics*, 82(2):80–89.

Izady, N. and Worthington, D. (2011). Setting Staffing Requirements for Time Dependent Queueing Networks : The Case of Accident and Emergency Departments. *European Journal of Operational Research*, pages 1–24.

Jalalpour, M., Gel, Y., and Levin, S. (2015). Forecasting demand for health services: Development of a publicly available toolbox. *Operations Research for Health Care*, 5:1–9.

Jennings, O. B., Mandelbaum, A., Massey, W. A., and Whitt, W. (1996). Staffing To Meet Demand. *Management Science*, 42(10):1383–1394.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, pages 338–354.

Konrad, R., DeSotto, K., Grocela, A., McAuley, P., Wang, J., Lyons, J., and Bruin, M. (2013). Modeling the impact of changing patient flow processes in an emergency department: Insights from a computer simulation study. *Operations Research for Health Care*, 2(4):66–74.

Littig, S. J. and Isken, M. W. (2007). Short term hospital occupancy prediction. *Health Care Management Science*, 10(1):47–66.

Massey, W. A. and Whitt, W. (1993). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems*, 13(1-3):183–250.

Mayhew, L. and Smith, D. (2008). Using queuing theory to analyse the Governments 4-h completion time target in Accident and Emergency departments. *Health Care Management Science*, 11(1):11–21.

Newell, G. (1966). The $M/G/\infty$ Queue. *SIAM Journal on Applied Mathematics*, 14(1):86–88.

Pidd, M. (2012). *Measuring the performance of public services: Principles and practice*. Cambridge University Press.

Proudlove, N., Black, S., and Fletcher, A. (2007). Or and the challenge to improve the nhs: modelling for insight and improvement in in-patient flows. *Journal of the Operational Research Society*, 58(2):145–158.

Proudlove, N. C., Gordon, K., and Boaden, R. (2003). Can good bed management solve the overcrowding in accident and emergency departments? *Emergency Medicine Journal*, 20(2):149–155.

Riordan, J. (1951). Telephone traffic time averages. *Bell System Technical Journal*, 30(4):1129–1144.

Royston, G. (2009). One hundred years of operational research in healthuk 1948–2048. *Journal of the Operational Research Society*, 60(1):S169–S179.

Tijms, H. C. (2003). *A first course in stochastic models*. John Wiley and Sons.

Utley, M., Gallivan, S., Pagel, C., and Richards, D. (2009). Analytical methods for calculating the distribution of the occupancy of each state within a multi-state flow system. *IMA Journal of Management Mathematics*, 20(4):345–355.

Utley, M., Gallivan, S., Treasure, T., and Valencia, O. (2003). Analytical methods for calculating the capacity required to operate an effective booked admissions policy for elective inpatient services. *Health Care Management Science*, 6(2):97–104.

Utley, M. and Worthington, D. (2012). Capacity planning. In *Handbook of Healthcare System Scheduling*, pages 11–30. Springer.

Weiss, N. A., Holmes, P. T., and Hardy, M. (2006). *A course in probability*. Pearson Addison Wesley.