Modelling Critical Care Unit Activities Through Queueing Theory

Izabela Komenda

School of Mathematics
Cardiff University

A thesis submitted for the degree of
Doctor of Philosophy

April 2013

Abstract

Critical Care Units (CCUs) are one of the most complex and expensive of all medical resources and hospital managers are challenged to meet the demand for critical care services with adequate capacity. The pressure on critical care beds is continuously increasing as new medical equipment provides the opportunity to save more patients lives. It is therefore crucial that beds are managed well and used efficiently. This thesis describes two major projects, the first undertaken in conjunction with the CCU at the University Hospital of Wales in Cardiff (UHW); and the second with two CCUs from the Aneurin Bevan Health Board.

In the first project data has been analysed to determine the flow of patients through the Unit. Admissions to CCUs were categorised under two headings: emergency, and elective. The length of stay in the CCU is heavily dependent on the admission category. In this thesis, both computer simulation and theoretical queueing models have been considered, which show how improvements in bed management may be achieved by considering these two categories of patients separately. The vast majority of previous literature in this field is concerned only with steady-state conditions, whereas in reality the processes are time-dependent. This thesis goes some way to addressing this deficiency.

The second project relates to work undertaken with managers from the Royal Gwent Hospital in Newport and at the Nevill Hall Hospital in Abergavenny. Data from both hospitals have been analysed to define arrival and service processes. A state-dependent theoretical queueing model has been considered which has been used to investigate the significance of combining the two units. The model has been also utilised to advise on the number of beds the new combined unit should have in order to satisfy targets quoted by the hospital managers.

In the final part of the thesis, consideration has been given to the impact of collaboration, or lack thereof, between hospitals using a game theoretical approach. The effect of patient diversion has been studied. To formally investigate the impact of patients transfers, a Markov chain model of the two CCUs has been set-up, each admitting two arrival streams: namely, their own patients and transfers from other hospital. Four different models were considered and for each model the effect of targets, demand and capacity were studied. The efficiency of a system which degrades due to selfish behaviour of its agents has been measured in terms of Price of Anarchy.

DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed .. . .
Date . . . . . .

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed .. . .
Date . . . . . .

STATEMENT 2

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged and explicit references given. A reference section is appended.

Signed .. . .
Date . . . . . .

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .. . .
Date . . . . . .

Acknowledgements

It gives me great pleasure in extending my gratitude to all those people who have supported me during the last three and half years of my research.

First and foremost, I would like to express my deepest appreciation to my supervisors, Professor Jeff Griffiths and Dr Vincent Knight, for their expertise, contributions, patience and support when they were most needed. One simply could not wish for better or friendlier supervisors.

My sincere thanks also go to Dr Mari Jones for guiding and enlightening me the first few months of my research.

I would like to express my gratitude to the Engineering and Physical Sciences Research Council for making this research financially possible.

I am extremely grateful to Andrew Nelson and Dr Martyn Read from University Hospital of Wales in Cardiff and Dr Stephen Dumont and Zoe Goodacre from the Aneurin Bevan Health Board for providing the data, help in increasing my understanding of Critical Care Unit performance and assistance throughout with anything I required.

I express my appreciation to my thesis examiners: Professor Steven Gallivan and Professor Paul Harper. Their thoughtful questions and comments were valued greatly.

In my work I have been blessed with the most friendly and cheerful group of fellow research students. I am reluctant to name individuals for fear of missing somebody out, but I would like to say thank you all for ensuring I kept level-headed during heated or stressed times - tea breaks and sweets normally did the trick! You will stay in my memories forever!

I would also like to acknowledge with much appreciation the staff of the School of Mathematics for welcoming me as a friend and providing the help, support and equipment I have needed.

I would like to express my sincere thanks to my family. In particular, I would like to extend a huge thank you to my Mum, who ensured I made the most of opportunities offered to me, who always believed in me and supported any decision I made in my life. Words fall short of her impact on my life. Without her, I would not be the person that I am today. Kocham Cię Mamo.

Last but not least, thank you God for giving me the intellect to understand the complexity of numbers, and giving me strength to complete this research.

Publications and Presentations

List of Publications

J. D. Griffiths, V. Knight, and I. Komenda. Bed management in a Critical Care Unit. *IMA Journal of Management Mathematics,*, 24(2): 137-153, January 2013. [74]

I. Komenda, J. D. Griffiths, and V. Knight. A model of CCU activities through queueing theory. *ORAHS Conference, 2012*, extended abstract only. [107]

Conference Contributions

I. Komenda, J. D. Griffiths, and M. Jones. Queue models applied to healthcare and transport. *SCOR Conference*, April 2010, Nottingham.

I. Komenda, J. D. Griffiths, and V. Knight. Bed management in the Critical Care Unit. *OR52 Conference*, September 2010, Royal Holloway University of London.

I. Komenda, J. D. Griffiths, and V. Knight. Bed management in the Critical Care Unit. *IFORS Conference*, July 2011, Melbourne, Australia.

I. Komenda, J. D. Griffiths, and V. Knight. Mathematical modelling of the critical care units at the Royal Gwent and Nevill Hall hospitals. *SCOR Conference*, April 2012, Nottingham.

I. Komenda, J. D. Griffiths, and V. Knight. Mathematical modelling of the critical care units at the Royal Gwent and Nevill Hall hospitals. *EURO Conference*, July 2012, Vilnius, Lithuania.

I. Komenda, J. D. Griffiths, and V. Knight. Mathematical modelling of the critical care units at the Royal Gwent and Nevill Hall hospitals. *ORAHS Conference*, July 2012, University of Twente, Holland.

Poster Presentation

I. Komenda, J. D. Griffiths, and V. Knight. How to improve bed management in a Critical Care Unit. *ORAHS Conference*, July 2011, Cardiff University, Wales.

Contents

List of Acronyms

A&E	Accident and Emergency
CART	Classification and Regression Tree
CCU	Critical Care Unit
CDF	Cumulative Distribution Function
CV	Coefficient of Variation
DoH	Department of Health
EAU	Emergency Assessment Unit
ED	Emergency Department
ER	Emergency Room
ENB	English National Board For Nursing
HDU	High Dependency Unit
ICNARC	The Intensive Care National Audit and Research Centre
IID	Independent and Identically Distributed
ICU	Intensive Care Unit
ITU	Intensive Treatment Unit
LoS	Length of stay
MAU	Medical Assessment Unit
NH	Nevill Hall Hospital in Abergavenny
NHS	National Health Service
OR	Operational Research
PDF	Probability Density Function
PoA	Price of Anarchy
RG	Royal Gwent Hospital in Newport
RIP	Riyadh ICU Program
SCCC	Specialist Critical Care Centre
UHW	University Hospital of Wales in Cardiff
VBA	Visual Basic for Applications

Chapter 1

Introduction

Operational Research (OR) is a relatively new sub-field of mathematics. It is thought to have been conceived in the efforts of military planners during World War II. The objective was to find the most effective utilisation of limited military resources by the use of quantitative techniques. In the years after the Second World War the application of OR moved toward more domestic concerns. By the early 1950s over forty Operational Research sections, which ranged in size and speciality, had been established in Great Britain (Goodeve, 1953 [61]). Many were based in the private sector whilst others could be found in government departments or research associations. Various applications of the discipline were studied; agriculture, civil aviation, the textile industry, property development, and healthcare. The paper entitled "Operational Research in Medicine" (Bailey, 1952 [10]) is perhaps the earliest publication that considers the application of OR in healthcare. Since then research within this field has been constantly increasing. Operational Research provides numerous methodologies and solution techniques for tackling healthcare problems such as: how many nurses should a hospital employ? How many beds should a hospital have to provide adequate care for all patients? OR offers a systematic approach to problem solving and allows for the characterisation of activities of an existing system using mathematical modelling.

1.1 Introduction to Critical Care Units

A Critical Care Unit (CCU), also sometimes known as an Intensive Therapy Unit (ITU) or Intensive Therapy Department, is a special ward that is found in most acute hospitals. It provides intensive care (treatment and monitoring) for people who are critically ill or are in an unstable condition. People in CCUs need constant medical support to keep their body functioning. They may not be able to breathe on their own and they have at least one organ failure. There are many different conditions and situations that can cause organ systems to fail. Some of the most common include: a severe accident, such as a road accident, a serious acute health condition, such as a heart attack or stroke, a severe infection, such as pneumonia or blood poisoning (sepsis). More importantly,

patients after major surgeries are also admitted into the CCU; this can either be a planned admission as part of recovery after surgery, or an emergency measure if there are complications during surgery. Medical equipment takes the place of failed organ functions while the person recovers. Patients who are able to breathe unaided and no longer need critical care will be transferred to a different ward to continue their recovery. The time it takes to recover completely varies greatly from person to person, and will also depend on a variety of factors such as age, overall level of health and fitness and the severity of the patients' condition.

1.1.1 Critical Care Unit Beds

The CCU beds are very expensive and a limited resource because they provide specialised monitoring equipment, a high degree of medical expertise and constant access to highly trained nurses. It was estimated by the Department of Health (DOH) in 2005-2006 that each CCU bed costs the National Health Service (NHS) around £1,800 a day, including the nursing cost (DOH2006, [129]). However, in 2006-2007 the DOH (DOH2007, [130]) changed their costing policy and now calculates the cost per CCU patient according to the number of organ failures they have rather than the average cost of a bed.

1.1.2 Critical Care Unit Nurses

Due to the severity of the illness of patients in the CCU, the general policy in the United Kingdom is to allocate one nurse to each critical care patient at all times. One nurse may care for two less sick patients, and occasionally a particularly sick patient may require two nurses. Elsewhere in Europe the nurse to patient ratio is usually 1:2 or 1:3, although the units are larger and have a higher proportion of low risk patients. Currently in the UHW CCU there are 24 nurses scheduled per shift. Many critical care nurses will have completed a specialist training programme and have extensive experience and expertise. Not surprisingly, nursing salaries comprise the largest component of the intensive care budget. It was estimated that a high percentage of CCU bed costs are the nursing costs ([129]).

However, a shortage of qualified staff exists, which leads to refused admissions, cancellation of major elective operations, and a heavy and stressful workload for the existing nurses. To ease this problem, healthcare assistants are being increasingly used to undertake some of the more routine tasks.

1.2 Problems that Critical Care Units are Facing

There are a few problems associated with the CCU. The first problem that the CCU has to deal with is the shortage of beds. On average, 8% of patients are refused admission to a CCU because the Unit is full (Audit Commission Report, 1999 [6]). The CCU occupancy rates for some hospitals are reportedly very high (Smith, 1995 [151] and Mitchell, 1995 [126]) and a shortage of beds has been identified throughout the UK.

Shortages of CCU beds can cause unwanted consequences to patients, some of which may prove fatal. The greatest impact of the insufficient number of CCU beds is on potential patients awaiting elective surgery. These patients are considered to be low priority patients. Major operations may be cancelled, often at a very short notice, because there is no available CCU bed in which the patient can recover post-operatively. In addition to the unnecessary stress generated by a surgery cancellation, the delay may have serious medical consequences. On average, the CCUs report three cancelled operations for every 100 patients who they were not able to admit (Audit Commission Report, 1999 [6]).

The second problem associated with critical care is a shortage of CCU trained staff. The severity of illness of critical care patients generates the need for a 1:1 nurse to patient ratio. Thus, the provision of more CCU beds would necessitate the employment of more CCU nursing staff. The specialist and high-skilled nature of critical care generates the requirement for a high proportion of nurses with relevant training. Many trusts have reported a distinct shortage of critical care nurses, particularly those trained to the English National Board For Nursing standard (ENB), which is the specialist training in critical care nursing (Audit Commission Report, 1999 [6]).

Further problems that CCUs have to deal with are costs. The cause and effects of shortages of resources, namely beds and nurses, were examined previously. An intuitive solution to these problems would be the provision of more resources. However, the relative benefits of providing more critical care beds and employing more nursing staff must be weighed up against the cost to the NHS. The provision of critical care is more expensive than other types of healthcare due to a higher staffing requirement, specialist equipment and therapeutic interventions. In fact, a study has shown that provision for a critical care patient can cost up to six times more than a patient on the general ward (Royal College of Anaesthetists and Royal College of Surgeons, 1996 [140]). The annual UK bill for critical care was estimated at £675-725 million in 1997 [46], with the conjecture that it is increasing at the rate of 5% each year. This increase in expenditure comes as a result of new interventions and medical advances, requiring more costly equipment. In addition, staff costs contribute to over 50% of the total CCU expenditure, with approximately 90% of this utilised to employ nurses ([6]).

One approach, based on mathematical modelling, that successfully addresses problems in the healthcare systems is queueing theory.

1.3 Queueing Theory

Queueing theory is one of many sub-disciplines of OR. Much of the initial work on queueing theory is attributed to Erlang. The author in his 1909 paper entitled "The Theory of Probabilities and Telephone Conversations" [47] analysed waiting times for call connections on the Copenhagen telephone system. The author continued his studies into the next decade and in 1917 published some of his most influential work. The paper "Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges" [48] established formulae for loss and waiting time - these have since become prominent results in the field.

Much of the theory is devoted to the derivation of performance measures evaluating characteristics such as the throughput, probability of delay, number of queueing items and the expected waiting time of customers in the queue (see, for example Stewart, 2009 [152]). Queueing theory may be utilised to ensure that queues do not build up excessively, whilst servers are active a reasonable proportion of time.

In the context of queueing theory, one may think of a service system as comprising of two elements: the service facility itself, which may be staffed by a number of servers; and a queue for service (except in specific cases where it may be specified that queueing is not permitted). At each facility, customers arrive and queue for some activity. Such a situation is depicted in Figure 1.1:

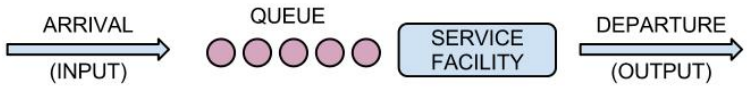


Figure 1.1: The fundamental diagram of queueing theory

Every queueing network is characterised by two major components: the arrival process and the service process. Queueing theory involves setting up mathematical models corresponding to Figure 1.1, analysing the system, and evaluating various performance measures. Since these processes are usually stochastic by nature, queueing theory is based on probabilistic analysis. The main characteristics are outlined below, and further details are given in Stewart, 2009 [152].

The arrival process defines how customers arrive at the service facility (e.g. singly or in batches) and how these arrivals are distributed in time. Throughout this thesis it is assumed, if not otherwise stated, that customers arrive at random with mean arrival rate λ, in a Poisson manner, so the inter-

arrival times are independently and exponentially distributed.

The queue discipline describes the order in which customers enter and leave the queue. This may be on a "first-in-first-out" ($FIFO$) basis, "first-in-last-out" ($FILO$) basis, "random-in-random-out" ($RIRO$) basis or in terms of priority.

The service mechanism outlines the resources needed for service to occur. The service time distribution defines how long the service will take, whilst other parameters such as the number of servers available, and whether the servers are in series (each server has a separate queue) or in parallel (one queue for all customers), must all be known in advance before analysis may be performed. In systems where the exponential distribution is assumed to provide an accurate representation of the distribution of service times, its Markovian (memoryless) property allows one to map the system to a continuous-time Markov chain which can be solved analytically. Similarly to the Poisson distribution, the exponential distribution is defined by a single parameter. To distinguish between the mean service time and the mean service rate, it is common to denote the mean service rate by μ, so that $\frac{1}{\mu}$ represents the mean service time.

Additional notation commonly used in the literature to analyse queueing systems, and that will be followed throughout this thesis, may be outlined as follows:

- P_n, $n = 0, 1, 2, \ldots$: the probability that there are n customers in the system
- c: the number of service channels

The quantity $\rho = \frac{\lambda}{c\mu}$, which is referred to as server utilisation rate, traffic intensity or load per server, is a common measure of interest that represents the behaviour of the queue over time. Essentially, if $\rho < 1$ then the servers are able to process customers faster than the rate at which they arrive, on average, so the queue will not grow infinitely long. If the system runs with $\rho < 1$ for an adequate period of time with stable mean service and inter-arrival rates, then all systems characteristics will eventually settle down and the system will run at a consistent level, considered as 'stable' or 'stationary'. When the system reaches this point of time, it is said to be operating in a steady-state fashion. It is this steady-state behaviour which has been intensively researched and is well-understood in the literature, since closed-form formulas have been derived to evaluate performance measures under these stationary conditions. The analysis of systems with non stationary arrival rates is however far more complex (Green *et al.*, 2006 [71]) and an overview of the literature on this topic is given in Section 4.2.1.

In 1953, a standard for the characterisation of queues was introduced by Kendall [98]. Kendall's notation provides a convenient classification of a queueing system in the form $A|B|C|D|E$ where:

(A) The inter-arrival distribution:

- M represents exponential, independent and identically distributed (IID) inter-arrival times;
- D represents deterministic (constant) IID inter-arrival times;
- E_k represents Erlang (with parameter k) IID inter-arrival times;
- G represents general IID inter-arrival times;

(B) The service time distribution (again commonly categorised as M, D, E_k or G. Phase-type distributions may also be used to specify systems with inter-related Poisson processes occuring within phases);

(C) The number of servers;

(D) The capacity of the system;

(E) The queue discipline e.g. $FIFO, FILO$ or $RIRO$;

For example in the $M|M|1|\infty|FIFO$ queueing system, M typically denotes Markovian inter-arrival / service times (i.e. exponentially distributed) and $FIFO$ denotes the queue discipline first-in first-out, with no restriction on the capacity of the system and one server.

Kendall's notation is commonly simplified to list only the first three characters $A|B|C$. In this format it is assumed that the queue discipline is $FIFO$, and no limits are imposed on the system capacity.

The $M/M/c$ model is one of the most widely researched models in the classic queueing literature since it is simultaneously capable of capturing randomness in arrival and service times. This permits the number of servers to be greater than one, and has the appealing benefit of a tractable steady-state solution. It represents a system with a single queue in which customers arrive at, and possibly queue, before being served by one of c servers. Arrivals occur according to a time-homogeneous Poisson process with a constant rate, and the service rate has an exponential distribution with a constant mean time. Such a system may be modelled as a basic birth-death process as described below (Stewart, 2009 [152]).

A birth-death process can be considered as a continuous time stochastic counting process $\{N(t), t \geq 0\}$. Letting $P_n(t) = \text{Prob}\{N(t) = n\}$ be the probability that the system is in state n at time t, the transition diagram of the birth-death process is illustrated in Figure 1.2. When a birth occurs, the system goes from state n to $n+1$ and when a death occurs it conversely goes from state n to $n-1$. The process is specified by birth rates $\{\lambda_i\}_{i=0,\ldots,\infty}$ and death rates $\{\mu_i\}_{i=1,\ldots,\infty}$

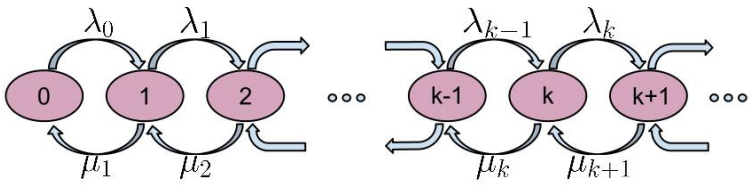


Figure 1.2: State diagram of a birth-death process

In queueing systems that directly model the behaviour of people, who arrive at a service facility requiring a specific service to be performed, the number of customers in the system is an appropriate state variable; $P_n(t)$ can be used to denote the probability that there are n customers in the system at time t. If λ_n does not depend on the number of customers in the system, then λ can be used to represent the mean arrival rate of customers. If μ represents the mean service rate provided at each of the c identical servers at all points in time, then $\mu_n = c\mu$ for $n \geq c$ and $\mu_n = n\mu$ for $1 \leq n < c$. Under these conditions, the state probabilities evolve according to the following differential-difference equations (Stewart 2009 [152]):

$$\begin{aligned}
\frac{dP_0(t)}{dt} &= -\lambda P_0(t) + \mu P_1(t), \\
\frac{dP_n(t)}{dt} &= -\lambda P_{n-1}(t) + (n+1)\mu P_{n+1}(t) - (\lambda + n\mu)P_n(t), && 1 \leq n < c, \\
\frac{dP_n(t)}{dt} &= -\lambda P_{n-1}(t) + c\mu P_{n+1}(t) - (\lambda + c\mu)P_n(t), && n \geq c.
\end{aligned} \tag{1.1}$$

Equation 1.1 is often referred to as the balance or Chapman-Kolmogorov forward differential equations. As the behaviour of the system settles to steady-state (as $t \to \infty$) then $P_0(t)$ and $P_n(t)$ are independent of time, so $\frac{dP_n(t)}{dt} = 0$ for $n = 0, 1, \ldots$, giving:

$$\begin{aligned}
&-\lambda P_0 + \mu P_1 = 0, \\
&\lambda P_{n-1} + (n+1)\mu P_{n+1} - (\lambda + n\mu)P_n = 0, && 1 \leq n < c, \\
&\lambda P_{n-1} + c\mu P_{n+1}(t) - (\lambda + c\mu)P_n, && n \geq c.
\end{aligned} \tag{1.2}$$

The steady-state probabilities defining the mean number of customers in the system are given by Equation 1.3 (for derivation of the summary measures, see Stewart 2009 [152]).

$$\begin{aligned}
P_0 &= \left[\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\mu^n} + \sum_{n=c}^{\infty} \frac{\lambda^n}{c^{n-c}c!\mu^n}\right]^{-1} \\
P_n &= \begin{cases} \frac{\lambda^n}{n!\mu^n} P_0, & \text{if } 1 \leq n \leq c-1 \\ \frac{\lambda^n}{c^{n-c}c!\mu^n} P_0, & \text{if } n \geq c. \end{cases}
\end{aligned} \tag{1.3}$$

The characteristics of $M/M/c$ systems permit relatively simple derivation of the number of cus-

tomers in the queue (L_q), the expected waiting time in the queue (W_q). While these measures both give insights into the degree of congestion that exists within a system, the distribution of the queueing time, and in particular the probability of waiting greater than time x in the queue $P(W_q > x)$ is often of greater interest, although more difficult to obtain analytically (Utley and Worthington, 2011 [158]).

Numerous authors such as Hershey *et al.*, 1981 [86] and Artalejo and Lopez-Herrero, 2001 [5] have since progressed Erlang's analysis of steady-state systems through deriving additional measures, including the moments of the length of a busy period and expected utilisation for constrained network facilities. In service systems governed by targets that specify minimum required standards, models can be set up to provide the performance of the system under various staffing levels and to find the minimum number of staff required to ensure the expected measures exceed the threshold levels (see Section 4.4). Yet, since the steady-state formulas are only capable of giving a single recommendation of an optimal nursing level (as they can only be applied to situations where the arrival of customers is strictly stationary), the earlier papers tend to place greater emphasis on system insights than the use of performance measures for this type of exploratory investigation.

Whilst much literature is devoted to the analysis of service systems with constant mean arrival and service rates (Green and Kolesar, 1991 [66]), most actual systems today are subject to time-varying demand, where arrival rates and the number of servers vary throughout the period of operation. Since admissions of elective patients to Critical Care Units are time dependent, Section 4.2.1 will provide more insight to time-dependent queueing theory.

1.4 Queueing Theory in Healthcare and Similar Environments

The use of queueing theory in a healthcare setting was rarely used until the pioneering work of Bailey, 1952 [10] appeared. In this paper queueing theory was used to develop an out-patient clinic scheduling system that gave acceptable results for patients (in terms of waiting time) and staff (in terms of utilisation). Homogeneity of patients was assumed as far as their service time distributions were concerned, and also it was assumed that all patients arrived for appointments on time. In more recent years, a vast number of queueing models have been developed for use in healthcare settings. Soon after, in 1954, a paper entitled "Queuing for Medical Care" was published by Bailey [11]. The author relates his study of an inpatient facility to Erlang's work on telephony by considering patients as telephone calls and hospital beds as telephone channels. The length of stay (LoS) is equivalent to the duration of the call. The author deduced the average waiting time (through Erlang's formula) and calculated the optimal number of beds required in the hospital. In subsequent years and decades, research interest in healthcare modelling through queueing theory has developed and there now exists a multitude of studies.

A considerable body of research has shown that queueing theory can be useful in real-world healthcare situations, and some reviews of this work have appeared. McClain, 1976 [121] reviewed research on models for evaluating the impact of bed assignment policies on utilisation, waiting time, and the probability of turning away patients. Nosek and Wilson, 2001 [131] reviewed the use of queueing theory in pharmacy applications with particular attention to improving customer satisfaction. Customer satisfaction is improved by predicting and reducing waiting times and adjusting staffing levels. Preater, 2002 [136] presented a brief history of the use of queueing theory in healthcare and points to an extensive bibliography of the research that lists many papers (however, it provides no description of the applications or results). Green, 2006 [65] presented the theory of queueing as applied in healthcare. The relationship between delays, utilisation and the number of servers was discussed, including the basic $M/M/c$ model, its assumptions and extensions, and the applications of the theory to determine the required number of servers. Fomundam and Herrmann, 2007 [53] summarised a range of queueing theory results in the following areas: waiting time and utilisation analysis, system design, and appointment systems. Their goal was to provide sufficient information to analysts who were interested in using queueing theory to model a healthcare process and who wanted to locate the details of relevant models.

The next section is an overview of research into using queueing theory as an analytical tool to predict how particular healthcare configurations affect delay in patient service and healthcare resource utilisation.

1.4.1 Waiting Time and Utilisation Analysis

In a queueing system, minimising the time that customers have to wait and maximising the utilisation of the servers or resources (doctors, nurses, hospital beds, etc.) are conflicting goals. This section is an overview of research into using queueing theory as an analytical tool to predict how particular healthcare configurations affect delay in patient service and healthcare resource utilisation.

1.4.1.1 Reneging

When a patient is waiting in a queue, they may decide to leave the system because they do not wish to wait any longer. Death on the waiting list is also an example of reneging. The probability that a patient reneges usually increases with the queue length and the patient's estimate of how long they must wait to be served. In systems where demand exceeds server capacity, reneging is the only way that a system attains a "state of dysfunctional equilibrium" (Hall *et al.*, 2006 [79]).
An important example of such a system is an emergency department. Broyles and Cochran, 2007 [19] calculated the percentage of patients who left an emergency department without getting help based on factors such as arrival rate, service rate, utilisation and capacity. From this percentage,

they determined the resulting revenue loss.

It is possible to redesign a queueing system to reduce reneging. A common approach is to separate patients by the type of service required. Roche *et al.*, 2007 [139] found that the number of patients who leave an emergency department, without being served, is reduced by separating out non-acute patients and treating them in dedicated fast-track areas. Most of their waiting would be for tests or test results after having first seen a doctor.

1.4.1.2 Bulking

Queueing system in which multiple customers arrive simultaneously and / or are served in groups of random size are called Bulk Queueing Systems (sometimes Batch Queueing Systems).

Batch service

The concept of batch service was introduced by Bailey, 1954 [11]. In his influential study, it was assumed that the inter-arrival rate followed the χ^2 distribution and utilised imbedded Markov chains to find the solution. He studied waiting times for an out-patient appointment with a hospital consultant and concluded that if a clinic was held once per week and the consultant was prepared to see at least one more patient than the average demand per week, then the average waiting time for an appointment would not be greater than one week.

The following year, Downton, 1955 [42] published results which were complementary to the work of Bailey. Dovnton calculated the variance of the waiting time distribution and was the first to publish tables of summary statistics for waiting times.

In 1976, Griffiths and Cresswell [73] applied the theory of batch service queues to a Pelican crossing. In this, the first application in this domain, several different aspects of the system mechanism were considered, such as the distribution of the queue of cars and pedestrians.

Chaudhry *et al.*, 1987 [25] considered the $M/G(a,b)/1$ queueing system. They employed the supplementary variable technique to find the post-departure and arbitrary time instant probabilities, and then used the results to find various summary measures including the mean queue length and the mean waiting time in the system.

Griffiths, 1995 [72] applied batch service queueing theory to the Suez Canal in a case study paper. The capacity of the Suez Canal was increased by 44%, by simply changing the cycle times from 24 to 48 hours.

Batch arrival

One of the first researchers to study batch arrivals was Donald Gaver, 1959 [59]. He considered the system whereby groups of customers arrived at a single service facility according to a stationary compound Poisson process and utilised embedded Markov chains to investigate the busy period.

Hawkes, 1965 [84] was the first to consider the time-dependent solution of a queue with bulk arrivals operating with a priority setting. Laplace transforms were used to consider a case for two classes of arrivals (priority and non-priority) and derived the equilibrium distribution (steady-state) for both classes of arrivals, as well as the distribution of the number of customers remaining in the system immediately after a departure occurred. The mean queueing times were also calculated for the case where the service time distribution was Negative Exponential.

One of the first explicit case study papers using the batch arrival queueing in a clinical setting was published by Lopezsoriano *et al.*, 1981 [114]. Different hospital departments released their staff at different times for their lunch, enforcing a batch arrival queueing scenario. The authors sought to optimise the system such that long queues and excessive waiting times for customers during the lunch period were minimised. 'What if' type scenarios were tested and the performance of the system was evaluated.

Jacob *et al.*, 1988 [94] discussed a queueing system with General inter-arrival and service times, one service facility and finite waiting space. The authors suggested two different rejection strategies which come into force when a batch arrives while there is not enough space in the buffer: either the entire batch is rejected or only the excess is rejected. The rejection probabilities were calculated and the waiting time distribution was considered.

Banik and Gupta, 2007 [12] investigated the system whereby customers arrived in batches to a finite buffer single server queue. The time between batch arrivals had a General distribution and the batch size was random. The service process was described as a Markovian service process. This model was then used to analyse two customer rejection policies, namely partial batch rejection and total batch rejection. Steady state distributions were developed at specific and arbitrary time instants. They obtained performance measures including the blocking probabilities and summary measures including waiting times.

1.4.1.3 Variable Arrival Rate

Although most analytical queueing models assume a constant customer arrival rate, many healthcare systems have a variable arrival rate. In some cases, the arrival rate may depend upon time but be independent of the system state. For instance, arrival rates change due to the time of day, the

day of the week, or the season of the year (see extended literature review in Section 4.2.1). In other cases, the arrival rate depends upon the state of the system (see Chapter 6).

Worthington, 1987 [169] presented an $M(\lambda q)/G/c$ model for service times of any fixed probability distribution and for arrival rates that decreased linearly with the queue length and the expected waiting time. The arrival rate may increase over time due to population growth or other factors. Rosenquist, 1987 [141] studied how an increase in patient arrival rate affected waiting times and queue length for an emergency radiology service.

A system with congestion discourages arrivals. Worthington, 1991 [170] suggested that increasing service capacity (the traditional method of attempting to reduce long queues) had little effect on queue length because as soon as patients realize that waiting times would reduce, the arrival rate increases, which increases the queue again.

1.4.1.4 Priority Queueing Discipline

In most healthcare settings, unless an appointment system is in place, the queue discipline is either first-in-first-out or a set of patient classes that have different priorities (as in Critical Care Units, which treats emergency patients with life-threatening injuries before elective patients).

Taylor *et al.*, 1969 [155] modelled an emergency anaesthetic department operating with priority queueing discipline. They were interested in the probability that a patient would have to wait more than a certain amount of time to be served.

Haussmann, 1970 [83] investigated the relationship between the composition of prioritized queues and the number of nurses responding to inpatient demands. The authors found that a slight increase in the number of patients assigned to a nurse and / or a patient mix with more high-priority demands resulted in very large waiting times for low priority patients.

McQuarrie, 1983 [123] showed that it is possible, when utilisation is high, to minimise waiting times by giving priority to clients who require shorter service times. This rule is a form of the shortest processing time rule that is known to minimise waiting times. It is found infrequently in practice due to the perceived unfairness (unless that class of customers is given a dedicated server, as in supermarket check-out systems) and the difficulty of estimating service times accurately.
Worthington, 1991 [170] analysed patient transfer from outpatient physicians to inpatient physicians. The patient was assigned one of three priority levels. Based on the priority level, there was a standard time period before which a referred patient should be scheduled to see the inpatient physician. The model assumed sufficient in-patient capacity to treat the highest priority category within

its standard time, and proposed sharing the remaining service capacity amongst the lower priority levels in such a manner that they each exceeded their standard target times by the same percentage.

Siddhartan *et al.*, 1996 [150] proposed a priority discipline for different categories of patients and then a first-in-first-out discipline for each category. They found that the priority discipline reduces the average wait time for all patients; however, while the wait time for higher priority patients reduced, lower priority patients endured a longer average waiting time.

Tuft and Gallivan, 2001 [157] used a computer simulation to compare three years' operation of different admission strategies: a first-come-first-served booking system, a triage booking system, and a waiting list system in which admissions were strictly ordered according to priority stratum. It was shown that the most effective system for minimising priority weighted delay is, at the time of outpatient assessment, to schedule surgery for the high priority patients for the first available operating slot, while assigning low priority patients to the most delayed slot that is feasible.

When arriving patients are placed in different queues, each of which has a different service priority, the queue discipline may be preemptive or non-preemptive. In the latter, low priority patients receive service only when no high priority patients are waiting, but the low priority patient who is receiving service is not interrupted if a high priority patient arrives and all servers are busy. In the preemptive queue discipline, however, the service to a low priority patient is interrupted in this event. Green, 2006 [65] presented models for both queue disciplines.

Fiems *et al.*, 2007 [52] investigated the effect of emergency requests on the waiting times of scheduled patients with deterministic processing times. It was a pre-emptive repeat priority queueing system in which the emergency patients interrupted the scheduled patients and the latter's service was restarted as opposed to being resumed. The authors modelled a single server queue and divided time into equally long slots. During periods when there is an emergency interruption, it was assumed that no server was available for non-emergency patients.

1.4.1.5 Blocking

Blocking occurs when a queueing system places a limit on queue length. For example, an outpatient clinic may turn away walk-in patients when its waiting room is full. In a Critical Care Unit, where patients can wait only in a bed, the limited number of beds may prevent a Unit from accepting patients.

Kabak, 1968 [95] was the first to consider c service facilities in this context when he developed the $M(n)/M/c$ batch arrival queue (with n arrivals in each batch). He examined the blocking prob-

abilities for a loss system and a delay system, and calculated the mean and variance of the delay time along with other numerical results.

McManus *et al.*, 2004 [122] presented a medical-surgical Intensive Care Unit where critically ill patients can not be put in a queue and had to be turned away when the facility was fully occupied. This is a special case, where the queue length can not be greater than zero, which is called a pure loss model (Green, 2006 [65]).

Koizumi *et al.*, 2005 [104] found that blocking in a chain of extended care, residential and assisted housing facilities resulted in upstream facilities holding patients longer than necessary. They analysed the effect of the capacity in downstream facilities on the queue lengths and waiting times of patients waiting to enter upstream facilities. System-wide congestion could be caused by bottlenecks at only one downstream facility.

Chydzinski and Winiarczyk, 2008 [28] considered the blocking probability in a finite-buffer queue with arrivals following a batch Markovian process (BMAP). Firstly the authors gave a comprehensive description of the BMAP under consideration. They then derived an expression for the transform of the blocking probability and demonstrated time-dependent and steady state characteristics from this expression. Numerical results were provided for two different types of BMAP.

More detailed literature review specific to bed blocking subject will be included in Sections 5.2.4 and 6.5.4.2.

The second approach that succesfully addresses problems in the healthcare systems is simulation.

1.5 Simulation

Many previous researchers have developed simulation and queueing models to help manage bed capacities in hospitals (Harper and Shahani, 2002 [82]; Gallivan and Utley, 2011 [57]; Dumas, 1984 [45]; Gorunescu *et al*, 2002 [62]; Cooper and Corcoran, 1974 [34]). The remainder of this discussion of relevant previous research will now focus in particular on simulation models developed specifically for CCUs.

Discrete event simulation has been widely utilised in modelling Intensive Care Units (ICUs); for example, Kim *et al*, 1999 [100] utilised a simulation model and queueing theory to describe activities in an ICU at a hospital in Hong Kong. Objectives of the initial study were to determine whether the ICU has sufficient capacity. The authors concluded that the current ICU capacity of 14 beds is sufficient to handle patients at the current arrival rates. Also, the reservation of some of the Unit's

beds for the sole use of elective patients was considered. Subsequent to this research, the simulation model was updated and used to evaluate methods of managing the existing beds more efficiently (Kim *et al*, 2000 [101]). It was found that the elective surgery patients caused the most disruption to the Unit and so a number of bed reservation schemes were evaluated. It was suggested that some ICU beds could be reserved for the exclusive use of elective surgery patients. It was proposed that this would reduce the number of cancelled elective surgeries and the simulation model supported this proposition.

The simulation model was utilised to explore the possibility of using elective surgery quotas (for example one per day) in conjunction with a scheduling window (one or two weeks) to reduce the demand fluctuations of patients requiring intensive care following elective surgical procedures and thus reducing elective surgery cancellation rates (Kim *et al*, 2002 [99]). It was determined that the combination of a daily quota schedule and reserving beds exclusively for elective surgery patients can greatly reduce the number of cancelled surgeries with minimal negative consequences for the other patients.

Classification and Regression Tree (CART) analysis is a very useful tool for the creation of similar patient groups and has been utilised in many simulation models. Shahani *et al*, 2008 [148] utilised CART analysis to create homogeneous groups of patients to feed into a simulation model. Several 'what if' scenarios were tested including an increase in capacity and the transfer of long stay patients onto a different ward.

Costa *et al*, 2003 [35] also used CART analysis to generate similar patient groups in the CCU. Their model gave emergency patients priority status over elective patients. Also they showed that capacity planning in the CCU can not simply be based on averages as this may generate an underestimation of resource needs during busy periods.

A dynamic simulation model was built of the CCU at the Cincinnati VA Medical Centre, Ohio by Cahill and Render, 1999 [22] to model the time varying behaviour of a system. They tested several alternative bed configurations to see whether a high bed occupancy level (81%) could be reduced to something more acceptable. It was found that the addition of telemetry and respiratory care beds would result in improved availability of ICU beds and that the addition of heart Emergency Room (ER) beds would resolve the ICU access problems. Unexpectedly, the increased ICU bed availability resulted in increased hospital bed utilisation and increased length of stay on the hospital service. It was therefore concluded that targeted reductions in length of stay would be needed before the implementation of the new plans.

In the Netherlands, Litvak *et al*, 2008 [113] constructed a model of several CCUs in the locality

and tested the scenario of reserving a pooled number of beds across the region for emergency admissions. The model could predict the optimal number of regional beds required for any given acceptance rate of emergency admissions.

The question of the maximum number of elective surgeries administered each day to avoid diversion or cancellations of surgeries was addressed by Kolker, 2008 [106]. The Unit under consideration was large (51 beds), and the optimal number of surgeries scheduled each day was deemed to be four.

Optimal nursing requirements were addressed in [75] by Griffiths *et al*, 2004. A discrete event simulation model of an ICU was built in *Simul8* and various 'what if' scenarios relating to nurse numbers were investigated.

1.6 Conclusions

Literature contained in this chapter has a general character and it highlights where and how queueing theory can be used in healthcare environment. A more detailed literature review specific to subjects described in this thesis is included in later sections:

- Simulation modelling of hospitals with special emphasis of modelling Critical Care Units (Section 1.5)
- Time-dependent aspects in queueing theory (Section 4.2.1)
- Staff requirements (Section 4.4.2)
- Resource planning and bed allocation (Section 6.5.2)
- Game theory (Section 7.2)

1.7 Outline and Structure of Thesis

The primary objective of this study is to show how a mathematical modelling approach is able to provide quantitative evidence to aid decision making in a critical care environment. Thus mathematical models of the Critical Care Unit (CCU) environment will be developed. Furthermore, consideration will be given to the impact of collaboration or lack thereof between hospitals using a game theoretical approach.

The research included in this thesis can be divided into three main parts as shown in Figure 1.3. The first part will carry out an analysis of data provided by the University Hospital of Wales in Cardiff (UHW) to determine arrival and service patterns, resource numbers and the flow of patients

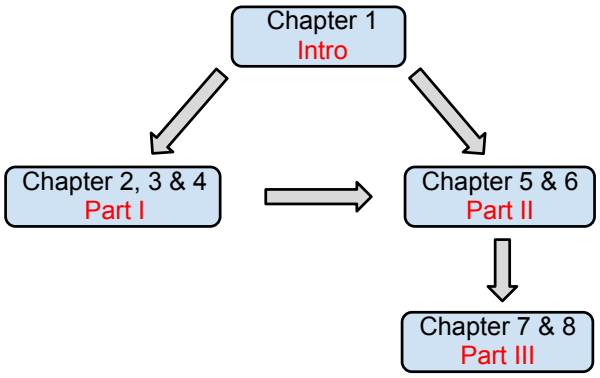


Figure 1.3: Thesis structure

through the Unit. Both computer simulation and theoretical queueing models will be considered. The models will be utilised to suggest improvements to the running of Unit, potential cost-saving measures, and the consequences of implementing certain new procedures such as admitting extra elective patients at non-busy times.

The second project describes work undertaken with managers from the Royal Gwent Hospital in Newport and at the Nevill Hall Hospital in Abergavenny. Data from both CCUs will be analysed to define arrival and service processes. A theoretical queueing model will be considered which will be used to investigate the significance of combining the two units. The model will be also utilised to advise on the number of beds the new combined unit should have in order to satisfy targets quoted by the hospital managers.

In the final part of the thesis, a game theoretical model will be proposed of two CCUs to study the effect of patient diversion. To investigate the impact of patients transfers, a Markov chain model of the two CCUs will be set-up, each admitting two arrival streams: namely, their own patients and transfers from other CCU.

Chapter 2

Summary Statistics of Patients' Flow Through the Critical Care Unit at the University Hospital of Wales

2.1 Introduction and Objective of the Study

It is intended that this chapter gives an insight of how the Critical Care Unit (CCU) at the University Hospital of Wales in Cardiff (UHW) operates. It will describe the importance of having critical care resources, including accessible beds and nurses available. The work described in Section 2.3 analyses the actual data from the CCU at the UHW to determine arrival and service patterns, resource numbers and the flow of patients through the Unit. Also, the main highlight of this chapter is that elective patients have very different profiles of admission and duration of stay from emergency patients; therefore any analysis should account for two different patient categories.

The primary objective of this project is to develop a mathematical model of the critical care environment. Computer simulation and theoretical models will be utilised to suggest improvements to the running of Unit, potential cost-saving measures, and the consequences of implementing new procedures.

2.2 Background

This part of the thesis investigates activities at the CCU at the UHW, which is the amalgamation of the previous ITU and HDU (High Dependency Unit). This amalgamation occurred in 2003 and the Unit has been running as a combined Unit ever since. The beds in the Unit can either be used as HDU or ITU beds. The CCU, which is the largest in Wales, consists of 24 beds with five additional beds that can be utilised in periods of peak demand. The five additional beds, which are based in

the Unit, are only utilised in very exceptional circumstances. These beds are currently unfunded, which means that to use any of them requires the employment of an agency nurse due to 1:1 or 1:2 nurse to patient ratio for most patients required in such a ward. Since it is the largest CCU in Wales, very often patients from all over Wales are treated at this Unit. The CCU beds are very expensive and a limited resource because they provide specialised monitoring equipment, a high degree of medical expertise and constant access to highly trained nurses.

2.2.1 Patients

The CCU at the UHW is the largest CCU in Wales it provides specialist care for a great percentage of the Welsh population. Patients are admitted onto the Unit from 6 different sources; Emergency Surgery, Elective Surgery, A&E, the Wards, Other Hospital and X-Ray.

Intensive care beds are occupied by patients with a wide range of clinical conditions, but all have a dysfunction or failure of at least one organ, particularly respiratory and cardiovascular systems. Patients usually require intensive monitoring, and most need some form of mechanical or pharmacological support such as mechanical ventilation or renal replacement therapy. As patients are admitted from most of departments in the hospital, staff in the CCU need to have a broad range of clinical experience and a holistic approach to patient care.

2.2.2 Data

The data set used in this study was provided by technical staff at the CCU of the UHW; it has fairly complete records for a period of six years; between 1^{st} of January 2004 and 31^{st} of December 2009.

As of June 1989, the collection of data at the UHW has followed guidelines proposed by the Riyadh ICU program (RIP), Medical Associated Software House Ltd, London, UK. The Riyadh predictive algorithm was first developed at the Riyadh Armed Forces Hospital, Saudi Arabia in 1984. The Riyadh ICU program formed the basis of the method by which many CCUs, including the CCU at the UHW, collate their data currently.

There are two main databases contained within the Riyadh ICU program. The first contains detailed patient information which is collected on the patients' admission to the Unit and on their departure. The second database contains data that is recorded on each day that the patient spends in the Unit.

Example of the information collected on admission to the CCU is: the patients' personal details (name, address, etc.). Note that personal details were not provided in our data set due to patient confidentiality; individual patients are identified using their CCU identification number. Other information collected are: patient demographics (age, gender, etc.), the source of the patient's arrival,

date and time of arrival and over a hundred of physiological factors detailing the patient's previous and current health status such as body temperature or blood pressure, measured on each day the patient stayed in the CCU. Data collected when the patient is discharged from the Unit includes date and time of discharge, whether or not they have survived, and the destination to which they are being discharged.

In the UHW there is a separate Paediatric Critical Care Unit, so only data on patients aged sixteen or over were used in this analysis. With these restrictions, and the condition regarding date of admission (each patient must be admitted on or before the 31^{st} December 2009) the data set available for use in the analysis has 8433 patients records.

2.3 Summary Statistics

The main objective of the work described in this section is to determine appropriate statistical distributions that could accurately represent the profile of arrivals to the CCU and lengths of stay in simulation and queueing models, which will be described in Chapter 3.

2.3.1 Arrivals and Discharges

Recall that patients are admitted to the CCU from six different sources. They are: Emergency Surgery, Elective Surgery, A&E, the Wards, Other Hospital and X-Ray. The percentages of patients who were referred from each source of admission are presented in Figure 2.1.

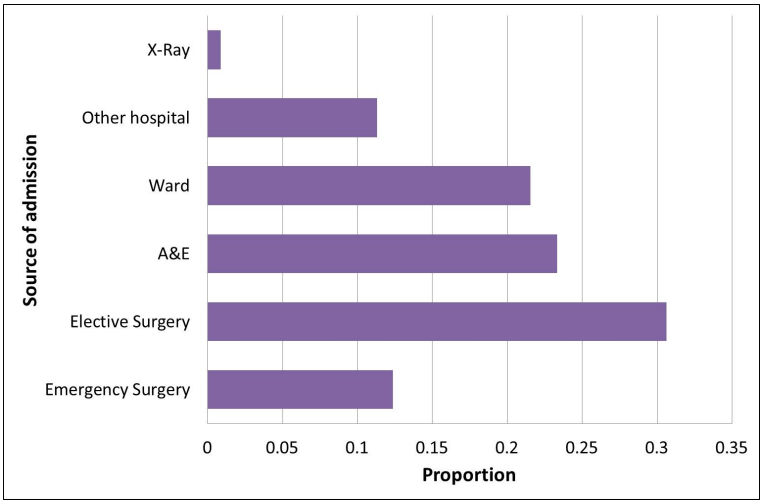


Figure 2.1: Source of admission

The largest source of admission is Elective Surgery which accounts for 31% of admissions. The smallest source of admittance is X-Ray which accounts for only 0.8% of all admissions.

While categorisation of patients by source of admission may be useful for hospital administration purposes, it is not entirely suitable from a modelling perspective. Categorising the patients according to whether or not their admission was planned would be much more revealing. Thus, for the remainder of this project the source of admission will be sub-classified into two groups; emergency and elective admissions. Emergency admissions are the unplanned admissions, over which the hospital has very little, or no, control. The second group are elective admissions, the planned admissions that hospital has control over.

It is not clearly apparent from the data set whether an admitted patient is an emergency or elective case, but some of the variables from the data set enable the patient type to by judged. The criterion that patients must satisfy to be classed as an elective patient are as follows. The surgery type must be elective, or post operative monitoring must have been planned. Also according to the Director of the CCU, 20% of patients admitted from wards, other hospitals or X-Ray following surgery are elective patients. Elective admissions account for 2687 cases, which is 31.86% of all admissions. The rest of the patients are emergency cases and they account for 5746 cases, which is 68.14% of all patients admitted.

The day of the week was considered for each admission. Table 2.1 indicates the percentage of arrivals on each day of the week during the study period (312 weeks). The Unit is facing the highest influx of patients on Thursday and Friday following by the lowest proportion of patients being admitted on Saturday and Sunday. To understand why this is the case, arrival data was split into emergency and elective category of arrival and subsequently into day of arrival (Monday-Sunday). The daily admittance patterns of patients by patient type are represented graphically in Figure 2.2.

Table 2.1: Percentage of admissions on each day of the week

Day of the week	**Percentage, %**
Monday	13.92
Tuesday	15.22
Wednesday	15.16
Thursday	17.31
Friday	16.36
Saturday	11.31
Sunday	10.72

Clearly, the number of elective arrivals is dependent on the day of the week. There is a large number of admissions on Wednesday, Thursday and Friday, with very few admissions on the weekend. This result is of course in accordance with expectations since surgeons would be required to work antisocial hours to operate on elective patients admitted on weekends. Also, from Monday through

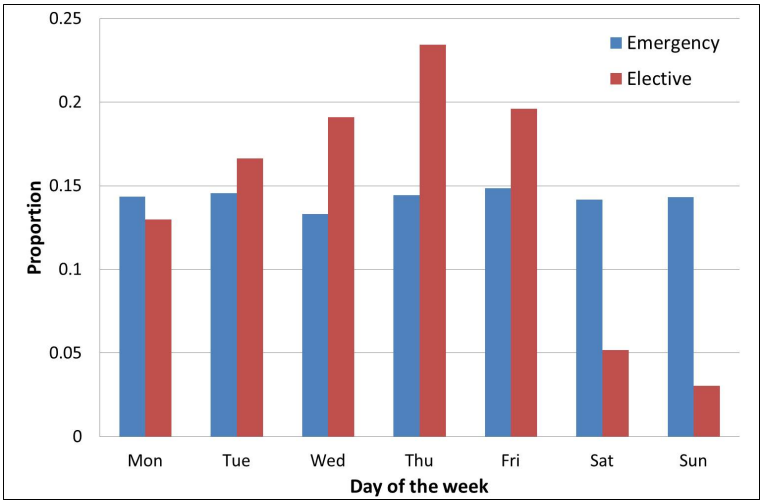


Figure 2.2: Daily arrival patterns for emergency and elective patients

to Friday, on average more than one elective arrival occurs each day.

By investigating emergency arrivals it can be seen that they are fairly equally distributed during the week. Recall that emergency patients are unplanned, possibly following a serious accident or rapid deterioration in condition. Therefore, distinct patterns in daily arrival times would not be expected for these patients.

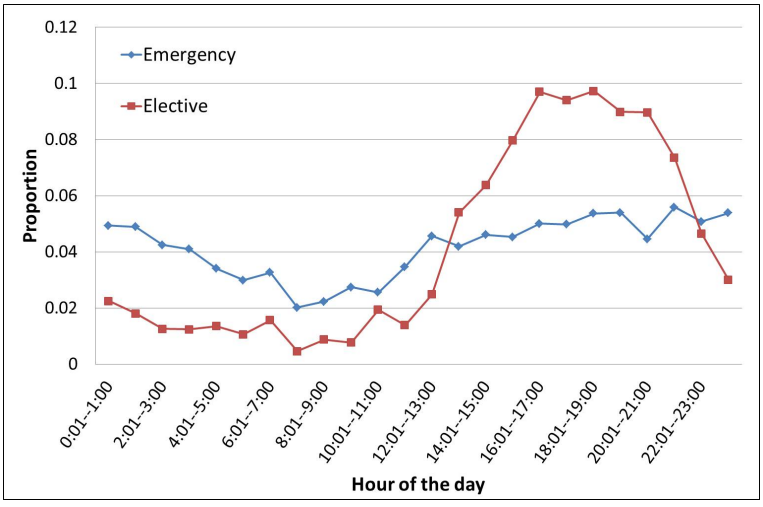


Figure 2.3: Hourly peaks of emergency and elective admissions

Any hourly admission trends are also examined. Figure 2.3 displays the percentage of admissions of emergency and elective patients, according to various hours of the day. There is an obvious hourly arrival trend for elective patients. The majority of admissions of patients following elective surgery are in the afternoon or evening, between 4pm and 9pm. The majority of surgeries are sched-

uled to start in the morning or afternoon and correspond to arrivals at the CCU in the afternoon or evening, depending on the duration of the surgical procedure. However, there is no obvious hourly trend for emergency patients except that there are very few admissions between 3am and 11am.

The arrival process needs to be analysed for later modelling purposes. Summary statistics for the daily arrival numbers at the CCU are given in Table 2.2. On average 3.83 patients get admitted to the CCU every day, but on some occasions data showed up to 12 admissions on one day.

Table 2.2: Summary statistics for the number of patients (elective plus emergency) admitted on each day

Summary statistic	**Value**
Mean	3.8312
Median	4
Standard Deviation	2.1136
Minimum	0
Maximum	12

On inspection of the frequencies of the number of admissions per day, it was suggested that the distribution may be modelled by a Poisson distribution. Analysis was carried out to investigate whether this was reasonable. The next task was to determine values for the parameter α which would provide the most appropriate fit to the distribution of the number of arrivals at the CCU each day. This was achieved using the optimisation tool: *Microsoft Excel Solver*. For the remaining of this thesis it will be called *Solver*.

Solver uses the simplex method to solve linear problems. The optimal solution is generated via an iterative method, where each solution has a lower objective function value than the previous. Hence, the solution becomes closer to the optimal after each iteration. However, care must be taken when searching for a global optimal solution, since local optima may be produced.

Consider now the formulation of the problem to be solved. Recall that values for the parameter α which would provide the most appropriate fit to the distribution of the number of arrivals at the CCU each day are required. The frequencies of the number of arrivals at the CCU on each day have been obtained from the data. Corresponding probabilities from the Poisson distribution were calculated by the probability distribution function (PDF):

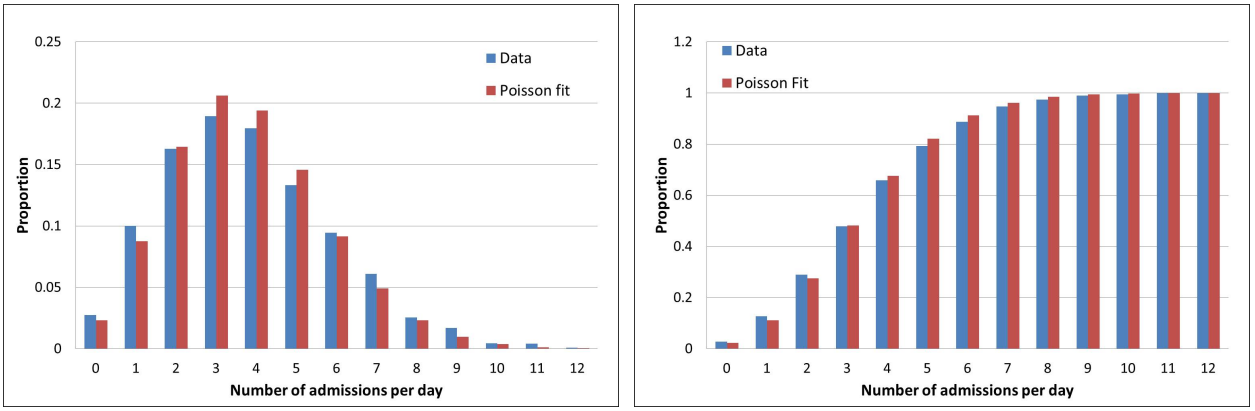
$$P(X = x) = \frac{e^{-\alpha}\alpha^{x}}{x!} \qquad\qquad x = 0, 1, 2, ...$$

Using an Excel spreadsheet, the deviations squared of these probabilities generated from the Pois-

son distribution and from the probabilities calculated from the actual data are evaluated. Finally, a calculation is made of the sum of the square of the deviations, and input into a single cell.

Solver is utilised to find the optimum values for α . The derived value of $\alpha = 3.7611$ gives a value for the sum of deviations squared of 0.00104, hence a good fit. In this case and for the remaining of this thesis no formal statistical tests will be performed to assess goodness of fits. The decision whether the distribution describes the data well will be based on graphical representation and the low value of the sum of deviations squared. An interesting observation was made by Raftery, 1995 [137], where he claimed that *P*-values and the tests based upon them give unsatisfactory results, especially in large samples.

Figure 2.4 displays a frequency of the number of admissions on each day along with the fitted Poisson distribution. Visibly, the Poisson distribution underestimates the frequencies of low numbers of admission and also high numbers of admissions, and consequently overestimates the frequency of a mid-number of admissions per day. Thus, the Poisson distribution in its present form is not quite suitable as a representation of the distribution of the number of daily arrivals to the CCU.



(a) PDF (b) CDF

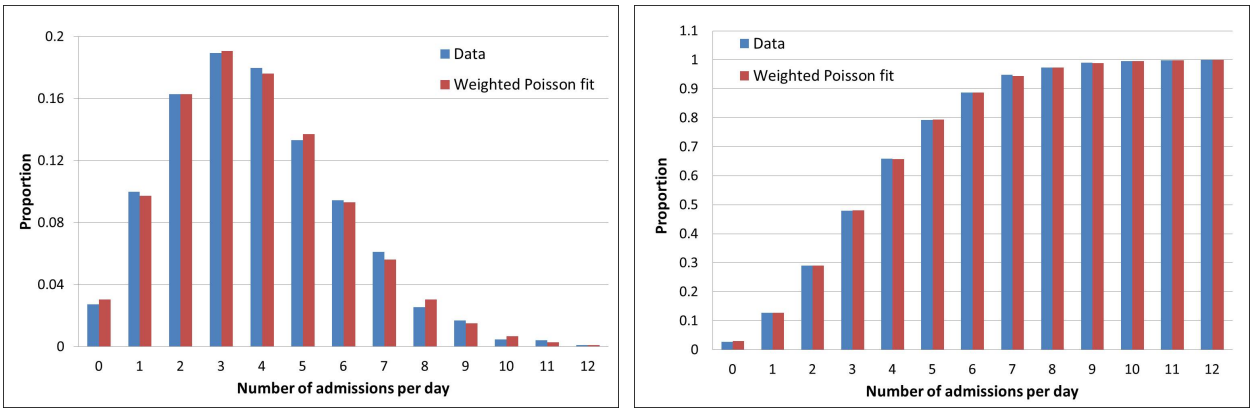
Figure 2.4: Poisson fit to the distribution of all admissions

It is decided to consider the Weighted Poisson Distribution, which could increase the probability of small and high number of admissions per day and decrease the probability of mid-number of arrivals per day. Corresponding probabilities of the number of arrivals from the Weighted Poisson distribution were calculated by the PDF:

$$P(X = x) = \omega \frac{e^{-\alpha_1} {\alpha_1}^x}{x!} + (1 - \omega) \frac{e^{-\alpha_2} {\alpha_2}^x}{x!} \qquad x = 0, 1, 2, ...$$

The Weighted Poisson distribution requires specification of three parameters, namely $\omega, \alpha_1, \alpha_2$ where ω must be the number between 0 and 1 and α_1, α_2 must both be positive. *Solver* is used

to minimise the sum of the squares of the deviation of the fitted distribution from the data. The corresponding optimum values are $\omega = 0.4518$, $\alpha_1 = 2.9022$, $\alpha_2 = 4.5958$, giving a value for the sum of the deviations squared of 0.00011. Figure 2.5 displays a frequency distribution of the number of admissions on each day along with the fitted Weighted Poisson distribution. The Weighted Poisson Distribution gives a much better fit than the standard Poisson Distribution.



(a) PDF

(b) CDF

Figure 2.5: Weighted Poisson fit to the distribution of all admissions

Admissions to the CCU appear to occur at random, with the exception of elective patients. On average 2.61 emergency patients and 1.22 elective patients are admitted every day. Since these means differ significantly, the arrival process needs to be considered separately for emergency and elective patients.

2.3.1.1 Emergency Admissions

Recall that distinct patterns in daily and hourly emergency arrival times are not noticeable. Table 2.3 gives the summary statistics for the number of emergency admissions per day.

Table 2.3: Summary statistics for the number of emergency admissions

Summary statistic	**Value**
Mean	2.6081
Median	2
Standard Deviation	1.6852
Minimum	0
Maximum	10

The Poisson arrival assumption has been shown to be a good one in studies of unscheduled arrivals (Young, 1965 [173]). However, on assessment of the frequencies, it is suggested that the presented

distribution again may be modelled by a Weighted Poisson distribution. Values for the parameters ω, α_1, α_2 which would provide the most appropriate fit to the distribution of the number of emergency arrivals at the CCU each day were determined using *Solver*. The optimum values are $\omega = 0.0333$, $\alpha_1 = 0.5918$, $\alpha_2 = 2.6540$, which give a value for the sum of deviations squared of 0.000049, confirming the goodness of fit. Figure 2.6 displays a frequency distribution of the number of emergency admissions on each day, along with the fitted Weighted Poisson distribution, which provides a very good fit.

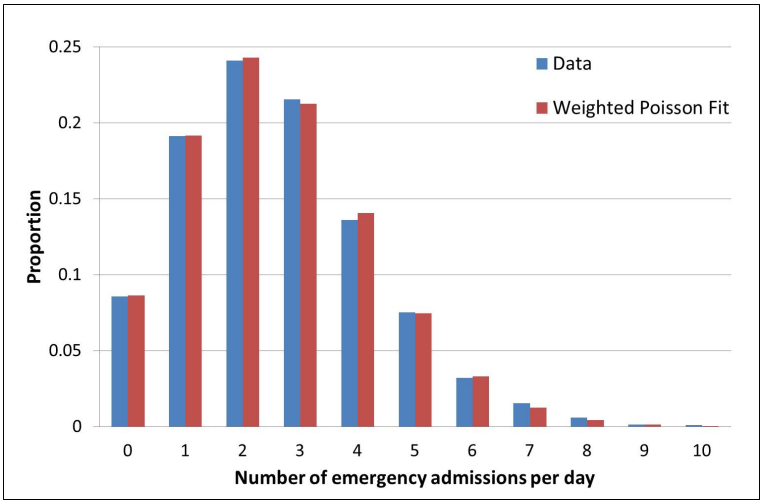


Figure 2.6: Weighted Poisson fit to the distribution of emergency admissions

2.3.1.2 Elective Admissions

Table 2.4 summarises statistical results for the number of elective admissions per day. On average 1.22 patients following elective surgery will be admitted every day for post-operative monitoring, but that number can rise to as high as seven.

Table 2.4: Summary statistics for the number of elective admissions

Summary statistic	**Value**
Mean	1.2231
Median	1
Standard Deviation	1.2689
Minimum	0
Maximum	7

On examination of the frequencies, it is suggested that the presented distribution may also be modelled by a Weighted Poisson distribution. The parameters are determined using *Solver*. The mini-

mum sum of deviations squared of 0.000025 is obtained for the values: $\omega = 0.3534$, $\alpha_1 = 0.3891$, $\alpha_2 = 1.6704$. The suggestion that a Weighted Poisson Distribution would give very good fit to the data is correct. Figure 2.7 displays a frequency distribution of daily elective admissions along with the fitted Weighted Poisson distribution, which evidently gives a near perfect fit.

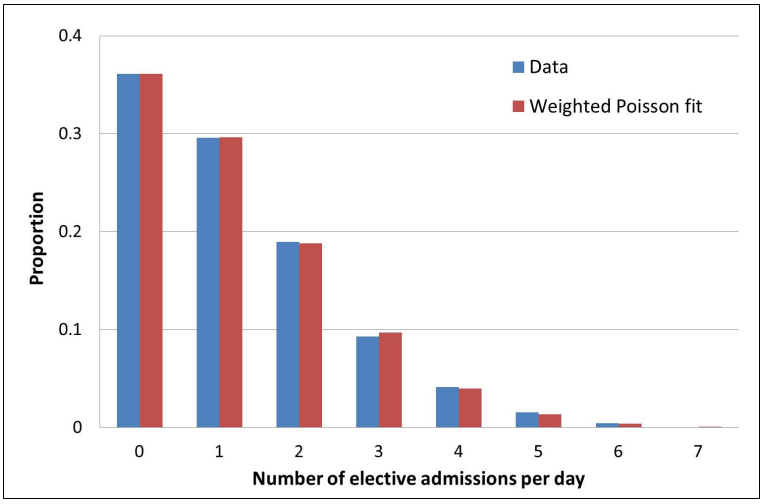


Figure 2.7: Weighted Poisson fit to the distribution of elective admissions

2.3.1.3 Discharges

The main objective of the work described in this section is to determine appropriate statistical distributions that could accurately represent the profile of discharges from the CCU. Patients are discharged from the CCU to four different destinations. They are: other wards, other hospitals, home or they are discharged as a result of death. Mortality in the CCU during the study period was on average 15.93% (the lowest mortality, 13.77%, being in year 2007, and highest, 18.94%, in year 2009) with approximately a further 6.1% dying on the ward after discharge from the CCU. Consider all discharges from the CCU. The percentages of patients' destinations are presented in Figure 2.8.

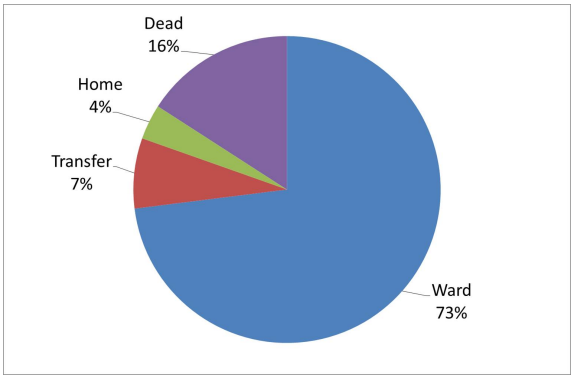


Figure 2.8: Post CCU destination

Alive patients can be discharged from the CCU for many reasons. The four main reasons are:

- The patient does not need any sort of life support and is fit enough to be discharged to a ward. 47.44% of all patients are discharged for that reason.
- The patient get transferred to a different hospital (6.65%).
- The patient or patient's family made a decision that they do not want to stay in the CCU any longer (16.74%).
- Existing critical care patients are sometimes discharged in order to accommodate higher priority patients (2.25%).

Patient's discharge could be delayed, because no beds were available on the destination ward (28.62%).

Similarly to admissions, discharges are considered by day of the week. Figure 2.9 indicates the percentage of discharges on each day of the week during the study period.

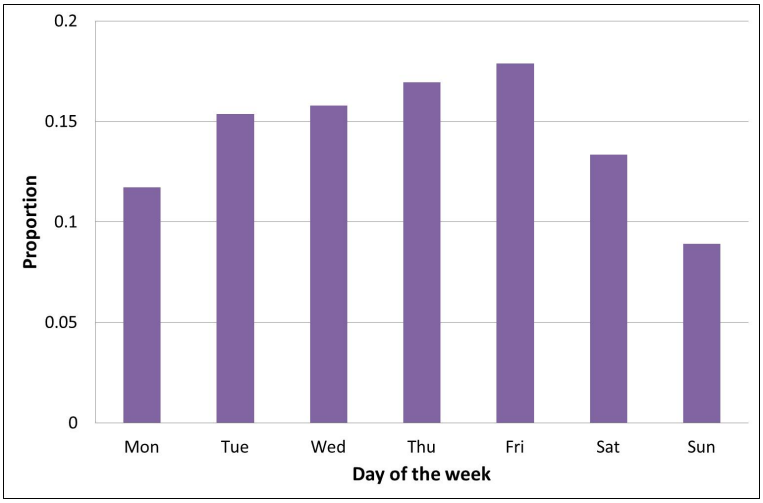


Figure 2.9: Daily discharge patterns

Visibly, there are fewer discharges on Saturday, Sunday and Monday, and the highest probability of discharge is on Friday. The reason could be that during the weekend there are fewer qualified clinicians in the Unit to decide whether the patient is fit enough to be discharged to a ward.

Any hourly discharge trends are now examined. In the data set the time of discharge was unavailable for 3 patients. It was assumed that these 3 patients were discharged at midday. Figure 2.10 displays the percentage of discharges, according to various hours of the day.

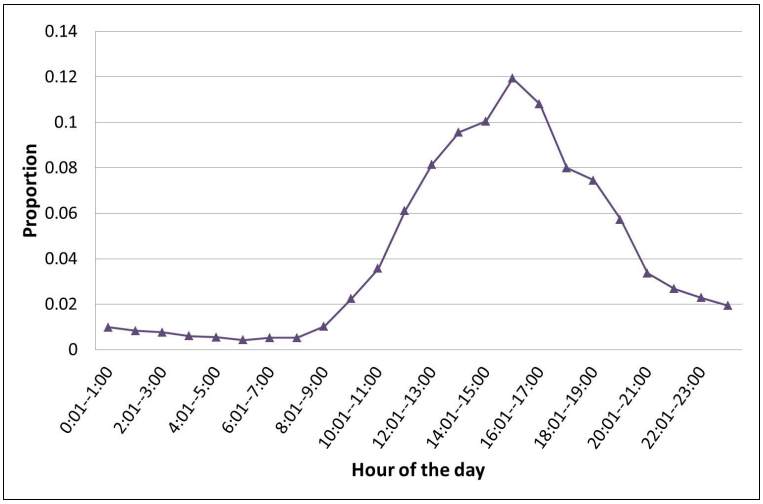


Figure 2.10: Hourly discharge patterns

There is obviously an hourly discharge trend. The majority of patients will be discharged in the afternoon between 2pm and 5pm and almost no discharges will appear in the night and early morning, between 11pm and 10am. Duke *et al*, 2004 [44] investigated the link between the time of discharge from CCU and mortality. Using Logistic regression, it was found that patients discharged during the night had a higher mortality rate.

As noted in Section 2.3.1, most of patients will be admitted between 5pm and 9pm; hence most discharges appear before 5pm in order to make a physical space for coming patients.

The number of patients that get discharged from the CCU each day was calculated; 24 patients discharged after the study period (01/01/2004-31/12/2009) were not taken into consideration. Table 2.5 presents the summary statistics for the number of patients discharged each day.

Table 2.5: Summary statistics for the number of discharges each day

Summary statistic	**Value**
Mean	3.8280
Median	4
Standard Deviation	2.1434
Minimum	0
Maximum	13

2.3.2 Inter-arrival Time

The time between consecutive arrivals is defined as an inter-arrival time. The data set contains time of admission, to the nearest minute, and the date of admission of each patient. After some manipulation, this enables the inter-arrival times to be calculated to the nearest minute.

Summary statistics of the emergency patient inter-arrival times are calculated and are displayed in the Table 2.6. The inter-arrival times are calculated to the nearest minute and then converted into hours for ease of presentation.

Table 2.6: Summary statistics for the emergency inter-arrival times

Summary statistic	**Value (hours)**
Mean	9.20
Median	6.08
Standard Deviation	9.53
Minimum	0
Maximum	78.75

The inter-arrival time distribution for emergency patients is modelled well using a Negative Exponential distribution. The result is not unexpected since arrivals often follow a Negative Exponential distribution (Coats, 2001 [30]) due to the random nature of emergency arrivals.

Figure 2.11 shows the comparison of the emergency inter-arrival distribution with the fitted Negative Exponential trendline.

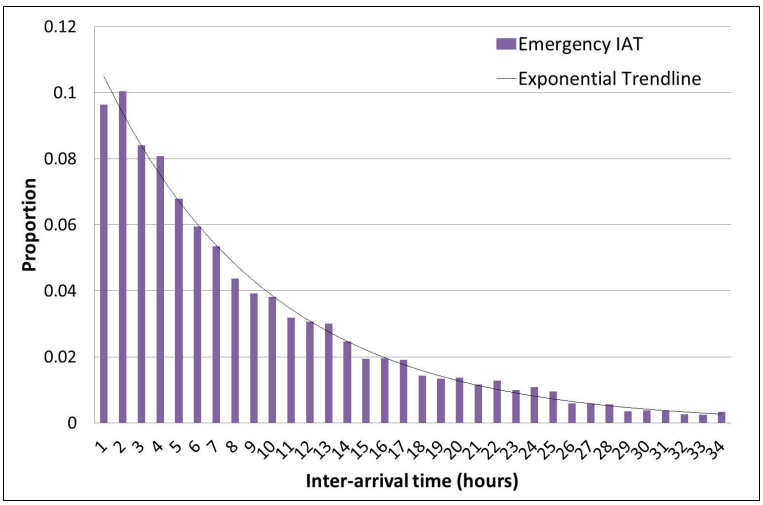


Figure 2.11: Emergency patients inter-arrival times

Next, consideration is given to the planned (elective) admissions; inter-arrival times are time-dependent in nature as visible in Figure 2.12.

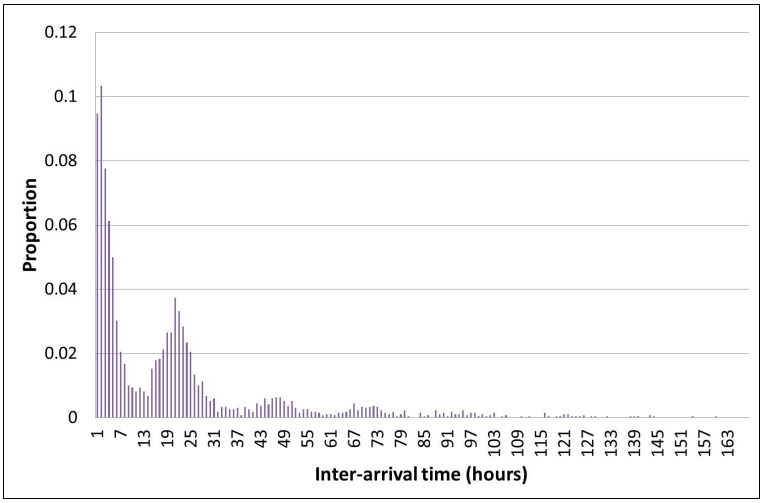


Figure 2.12: Elective patients inter-arrival times

Figure 2.12 shows that for many elective arrivals the inter-arrival time is between zero and four hours. Another peak occurs between 22 and 26 hour, and then other smaller peaks at approximately 24 hour intervals. These observations are consistent with the practice of elective surgery being performed at specific sessions spaced at 24 hour intervals. Inspection of the hour at which each elective patient arrives to the CCU showed that the majority of arrivals occur in the early evening with very few in the early hours of the morning.

Summary statistics of the elective patient inter-arrival times are calculated and are displayed in Table 2.7.

Table 2.7: Summary statistics of the elective patient inter-arrival times

Summary statistic	**Value (hours)**
Mean	19.62
Median	13.17
Standard Deviation	24.73
Minimum	0
Maximum	210

2.3.3 Length of stay

Length of stay (LoS) in CCU varies widely (Gallivan *et al.*, 2002 [58]). Some patients may require support for several weeks or months. These patients often have multiple organ failure. Patients

whose surgery had been planned are likely to experience a shorter recovery period with more positive outcomes. Most elective patients are discharged within 1-2 days, after post-operative monitoring. The outcome of patients who required emergency surgery is less predictable.

It is necessary to obtain more detailed observations to get an appropriate distribution of LoS. This is possible by making use of information contained within the data set. The data set provides information regarding the dates and times of arrival and discharge of each patient. This enables the LoS of each patient to be calculated, accurate to the nearest minute. In order to accurately calculate duration of stay, date of admission and discharge first need to be converted to an integer value using Excel function *datevalue* which converts a date in the form of text to a number that represents the date in Excel. Secondly, time of admission and discharge also needs to be converted to a number between 0 and 0.999988426 (0 is 12:00:00AM and 0.999988426 is 11:59:59) using Excel function *timevalue*, which converts a text time to an Excel serial number for a time. The duration of time spent in the CCU for each patient, in minutes, is calculated by subtracting converted admission time from converted discharge time. For ease of presentation, these values are converted into days, accurate to four decimal places.

The principal objective of this section is to determine a distribution that may accurately generate the duration of time that each patient spends receiving critical care therapy. Consider the frequency distribution of LoS of all patients in the CCU, presented in Figure 2.13. Each bar represents a two day period. 148 observations with LoS greater than 40 days are excluded from the graph for presentation purposes. Highly skewed LoS distributions have been observed in the literature previously (Faddy and McClean, 1999 [50]; Gorunescu, 2002 [62] and Marshal and McClean, 2003 [118]).

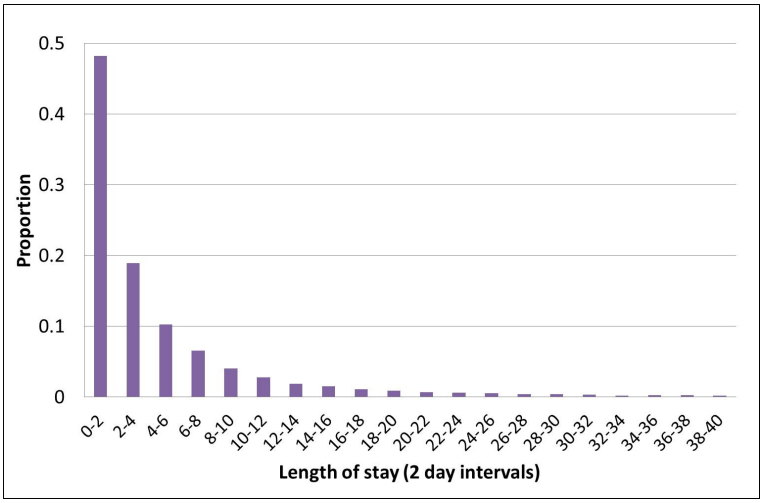


Figure 2.13: Length of stay frequency distribution

The data contains a vast range of values (from a minimum of 10 minutes to a maximum of 557 days). A number of relevant summary statistics of the data are considered; these are presented in Table 2.8.

Table 2.8: Summary statistics of LoS (days)

Summary statistic	**Value (days)**
Mean	5.8227
Median	2.1875
Standard Deviation	15.1485
Minimum	0.0069 (approx. 10 minutes)
Maximum	557.64

On assessment of the frequencies displayed and further analysis of the coefficient of variation (CV), which is a normalized measure of dispersion of a probability distribution, it is deduced that the Negative Exponential would not provide a very good fit. The coefficient of variation is defined as the ratio of the standard deviation to the mean, and for a Negative Exponential distribution the coefficient of variation is 1 or very close to 1. Clearly, $CV = \frac{15.1485}{5.8227} = 2.6016$ is greater than that of an Exponential distribution, suggesting that a Weighted Negative Exponential Distribution might give a better fit.

A Weighted Exponential, also known as a two-term mixed-exponential (Gorunescu *et al.*, 2002 [62]), has the probability distribution function of the random variable x as follows:

$$f(x) = \omega\beta_1 e^{-\beta_1 x} + (1-\omega)\beta_2 e^{-\beta_2 x}, \qquad x > 0 \tag{2.1}$$

This distribution often deals adequately with data which has a significant probability of obtaining a small value and a small probability of obtaining a large value (i.e. long tails).

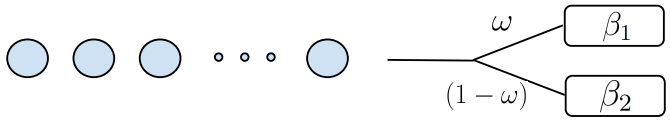


Figure 2.14: A Weighted Negative Exponential queueing system

A pictorial representation of a Weighted Negative Exponential distribution is illustrated in Figure 2.14. Customers arrive singly and form a queue. With a probability ω, they enter the top branch of the service facility where they will be served according to the Negative Exponential distribution with a mean rate β_1. With a probability $(1-\omega)$ they will enter the bottom branch of service

where they will be served according to the Negative Exponential distribution with a mean rate β_2. The service rate for each branch has been chosen in order to ensure that the overall service rate is $\beta = \frac{1}{\text{mean LoS}}$.

The queueing system demonstrated in Figure 2.14 looks very similar to the Hyper-exponential system with two phases; however, the mean service rates are different (as shown in Figure 2.15). Customers enter the top branch of the service facility with probability ω and they will be served according to the Negative Exponential distribution with a mean rate $2\beta\omega$. They will enter the bottom branch of service with probability $(1-\omega)$ where they will be served according to the Negative Exponential distribution with a mean rate $2\beta(1-\omega)$. The service rates are influenced by the probabilities that each branch is chosen, i.e. they both contain an element ω.

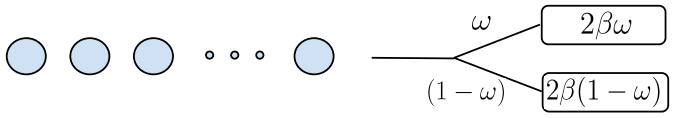


Figure 2.15: The Hyper-exponential queueing system with two phases

The probability distribution function of the random variable x having a two phase Hyper-exponential distribution is:

$$f(x) = \omega(2\mu\omega)e^{-2\mu\omega x} + (1-\omega)\left(2\mu(1-\omega)\right)e^{-2\mu(1-\omega)x}, \qquad x > 0 \tag{2.2}$$

The Weighted Negative Exponential provides a better fit than the Hyper-exponential distribution. The Weighted Negative Exponential distribution requires specification of three parameters, namely ω, β_1, β_2 where ω must be a number between 0 and 1 and β_1, β_2 must both be positive. The Hyper-exponential distribution is less flexible as it has only two parameters ω and β. The next task is to determine values for the parameters which would provide the most appropriate fit to the distribution of the length of stay in the CCU. This is achieved using *Solver*.

Consider now the formulation of the problem to be solved. The frequencies of the length of stay in the CCU for 0-2 days, 2-4 days, etc. were obtained from the actual data. Corresponding probabilities from the Weighted Negative Exponential distribution are calculated by integrating the probability distribution function (Equation 2.1) between the relevant limits (0-2, 2-4, 4-6, etc.), considering any suitable values for the parameters of the distribution at this stage. Using an Excel spreadsheet, the deviations of these probabilities generated from the Weighted Negative Exponential distribution from the data are calculated. Finally, the sum of the squares of the deviations are calculated, and input into a single cell.

Solver is utilised and the determined parameter values are: $\omega = 0.5528$, $\beta_1 = 0.1549$, $\beta_2 = 0.6935$, which give a value for the sum of deviations squared of 0.000098. The Weighted Negative Exponential distribution with the aforementioned parameters provides a near perfect fit to the distribution of the length of stay in the CCU.

Although LoS data is a continuous variable, the simplest method for comparing the data with the fitted distribution is to plot their respective frequency distribution on a single axis. Consider Figure 2.16, which presents the distribution of the length of stay in the CCU against the Weighted Negative Exponential distribution.

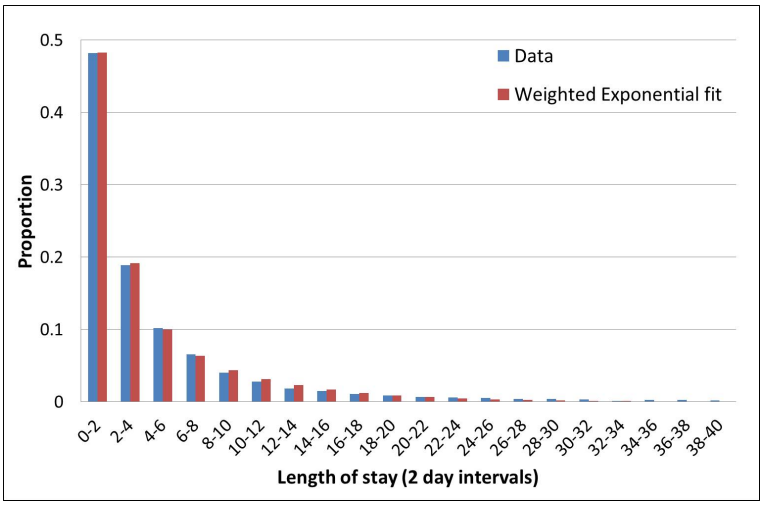


Figure 2.16: Length of stay frequencies with fitted Weighted Negative Exponential distribution

The Weighted Exponential distribution provides a very good fit to the LoS distribution. This is a significant contribution to the study of CCU LoS distributions since a number of authors (see, e.g. Harper and Shahani, 2002 [82]) have commented on the difficulty of selecting an appropriate theoretical distribution to match CCU LoS observations. The physical interpretation of the two-branch layout, inherent in the weighted-exponential systems, could well be that acute patients would be observed in one branch, while longer stay patients would move through the other branch.

Recall that the expected length of stay in the CCU is dependent on the patients' type. Thus, emergency patients and elective patients length of stay will now be considered separately.

2.3.3.1 Emergency Patients

Emergency patients are admitted to the CCU following a severe accident or unforeseen complication. The recovery period for those patients in comparison with those whose surgery is planned is expected to be longer. Also, the number of emergency patients who stayed in the CCU for a long period of time (longer than 40 days) is much higher than the number of elective patients (133

emergency patients compared to 15 elective patients).

Consider some summary statistics of the data regarding the length of stay that emergency patients spent in the CCU.

Table 2.9: Summary statistics of the length of stay (days) for emergency patients

Summary statistic	**Value (days)**
Mean	7.1109
Median	3.0468
Standard Deviation	17.6808
Minimum	0.0069
Maximum	557

The distribution of the length of stay that emergency patients spent in the CCU is now considered. As previously suggested, the Weighted Negative Exponential distribution might provide a good fit. Using *Solver* the optimum values of $\omega = 0.3381, \mu_1 = 0.0929, \mu_2 = 0.3366$ are obtained and they provide a very good fit to the distribution of the length of stay in the CCU for emergency patients. The results in Figure 2.17 show a good fit to the data which the Weighted Negative Exponential distribution gives. For most of the groups, the theoretical fit compares favourably with the actual data.

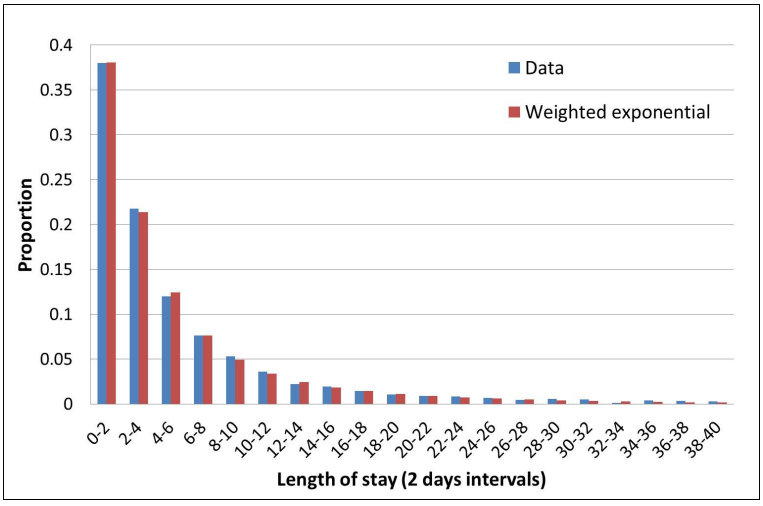


Figure 2.17: Length of stay distribution with data and the Weighted Exponential fit for emergency patients

2.3.3.2 Elective Patients

Patients admitted to the CCU following elective surgery are usually expected to make a quick recovery. The mean LoS of patients in this category is 3.06 days compared with the overall average of 5.81 days. Summary statistics of LoS in the CCU for these patients is presented in Table 2.10

Table 2.10: Summary statistics of the length of stay (days) for elective patients

Summary statistic	**Value (days)**
Mean	3.0678
Median	1.0208
Standard Deviation	6.3725
Minimum	0.0243 (approx. 35 minutes)
Maximum	74.5520

Consider now the distribution of the length of stay that a patient recovering from elective surgery spends in the CCU. As before, *Solver* is used to determine values for the required parameters. They are $\omega = 0.4452$, $\beta_1 = 0.2415$, $\beta_2 = 1.5062$, which give a value for the sum of deviations squared of 0.0002. Again, the length of stay for elective patients is obtained from the Weighted Negative distribution and results, together with the data are presented in Figure 2.18. For most of the groups, the observed values compare favourably with the theoretical ones.

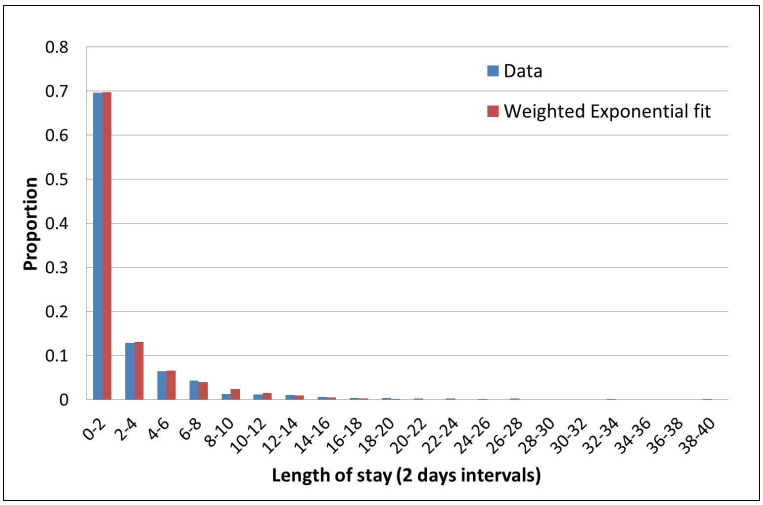


Figure 2.18: Length of stay frequencies with the Weighted Exponential fit for elective patients

Further analysis on whether the day of the week when the patient is admitted to the CCU has an influence on elective patients' length of stay is undertaken. It is concluded that if a patient is admitted to the CCU on Saturday or Sunday their length of stay will be on average 2 days longer comparing the length of stay of elective patient admitted on a weekday. In general, very few

elective admissions take place at the weekends suggesting that their condition is very serious and those patients very likely will require longer hospitalisation.

2.3.4 Bed Occupancy

The measure of bed occupancy is not provided directly by the Riyadh CCU Program, but can be evaluated by examining the admission and discharge times. An initial daily bed occupancy profile is developed using the date of the patients' arrival and date of discharge. However, it became obvious that this method generated an overestimation in the actual bed occupancy. Consider the situation, where a patient is discharged from the CCU at 10am, leaving an empty bed. If the next admission is at 4pm the same bed might be occupied. In this situation the number of beds occupied for that day would be recorded as two even though this is clearly not the case. Thus a different method of calculating bed occupancy is proposed.

A program written in Visual Basic reads in the number of patients that were in the Unit every hour from 01/04/2004 to 31/12/2009 and bed occupancy just before midnight is output to an Excel worksheet. The reason for omitting a period of three months (01/01/2004 to 31/03/2004) is to avoid underestimation; patients could have been admitted for example on 31/12/2003 and stayed in the Unit for a month. Records of that patient would not be included in the data set and therefore the fact that the patient occupied a bed would have been skipped. The number of beds occupied on each day from 01/04/2004 to 31/12/2009 is presented in Figure 2.19.

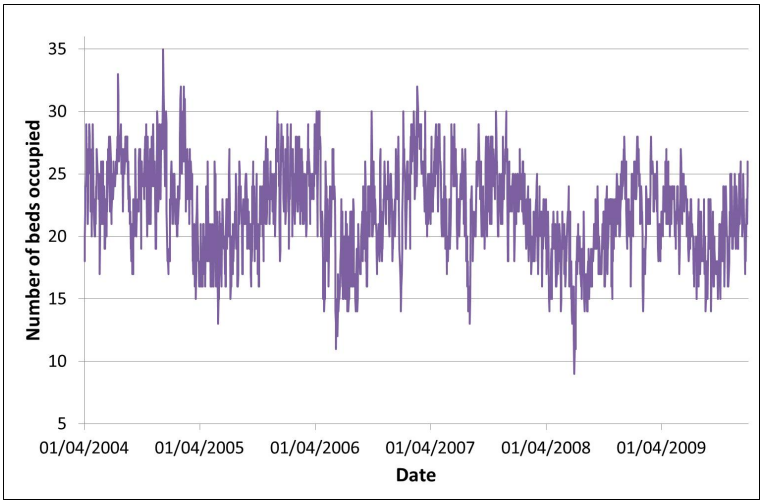


Figure 2.19: Midnight bed occupancy from January 2004 to December 2009

The mean number of beds occupied in the study period is 21.85 and standard deviation 3.65. Visibly, the number of beds occupied fluctuates dramatically, rising to 35 during busy periods and dropping to as low as 9 at quiet times. The CCU has 24 beds with an additional five available for

use at times of high demand. Initial analysis revealed that on 24% of occasions there were more than 24 beds occupied and on 1.32% of occasions more than 29 beds occupied. On this 1.32% of occasions, patients were most likely admitted onto so called 'virtual beds' in the Recovery Room or in the Cardiac Intensive Care Unit. Also, another reason for that high bed occupancy could be that when a patient dies it takes time for the nurses to prepare that bed for the next patient, which means that patient would feasibly be waiting on a trolley for a period of time.

The weekly trends of bed occupancy are investigated and results are presented in graphical form in Figure 2.20. Evidently, there is a pattern; the weekend bed occupancy is lower than during the weekdays. Seasonal trends are also investigated; it appears that the CCU experiences an influx of patients during the winter months (January, February, March) and have on average the highest bed occupancy (22.86 beds occupied). During the summer months (July, August, September) the CCU is not expected to experience such high bed occupancies; on average 20.62 beds are occupied during the summer, compared with the spring, 21.34 beds and the autumn 22.63 beds.

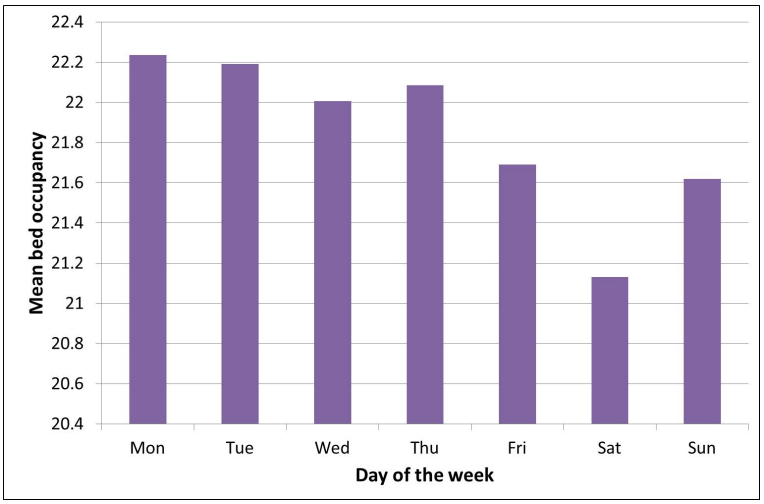


Figure 2.20: Bed occupancy depending on the day of week

Year 2007 appeared to be the busiest year; the average number of occupied beds was 23.78, which is 1.93 beds more than the overall mean. The next year, 2008, was the quietest; only 19.88 beds were occupied on average.

Consider the frequency distribution of the bed occupancy in the CCU during the study period, presented in Figure 2.21.

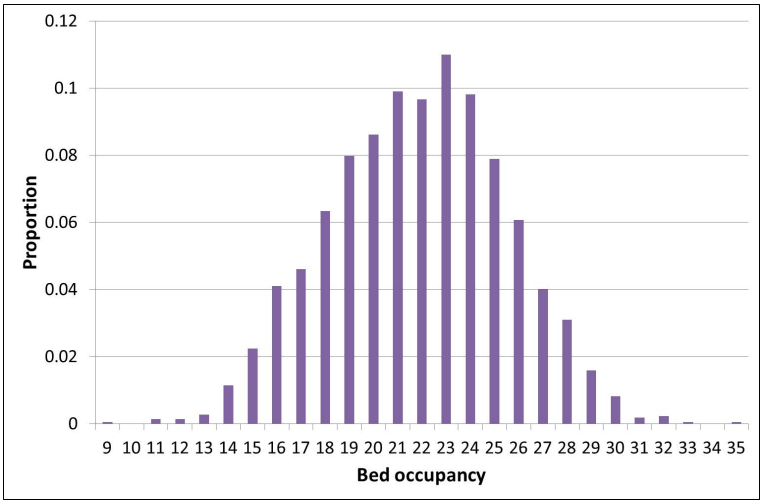


Figure 2.21: Bed occupancy frequency distribution

One of the main difficulties which the Director of the CCU faces lies in the unpredictability of the number of beds which will be occupied on any particular day. Figure 2.19 and 2.21 show a large degree of variation in daily bed occupancy over a five years and 8 months period. As was noted before, daily occupancy can be as low as 9, and as high as 35. This high degree of variability has many important implications. For example:

- If all beds are occupied when an emergency admission occurs, then that patient's condition may worsen while action is taken to find a bed, possibly in another hospital some distance away.
- If the number of nurses employed on a particular shift is insufficient, then agency nurses have to be employed at a cost of about three times that of the hospital's own nurses.

2.4 Conclusions

The main highlight of this chapter is that elective patients have very different profiles of admission and duration of stay from emergency patients; therefore any analysis accounted for two different patient categories. This chapter provided adequate fits to data which describe arrival processes and durations of stay in the Unit for each patient category. It has been shown that the Weighted Poisson distribution provided better fit for the daily number of arrivals than the Poisson distribution. Also, the Weighted Negative Exponential provided closer fit than the Negative Exponential to the duration of stay in the CCU.

Information included in this chapter will be used in the next chapter to consider aspects of theoretical and practical applications of mathematical modelling of the CCU.

Chapter 3

Mathematical Modelling of the Critical Care Unit at the University Hospital of Wales

3.1 Introduction

Mathematical modelling is a tool used to investigate situations whereby actual experiments would be impractical; for example, the modelling of future events or situations where cost would be an issue. Both simulation modelling and theoretical queueing techniques will be utilised in this chapter, which can help to improve the unit's capacity utilisation at relatively low cost and at little risk.

Two main aims of this chapter are to suggest measures which may be implemented to increase the throughput of patients in the Critical Care Unit (CCU) at the University Hospital of Wales (UHW), and to determine ways in which the degree of variation in daily bed occupancy may be reduced.

3.2 Simulation Model of the Critical Care Unit

The variability of the CCU environment must be considered when planning and managing resources. The development of a simulation model of the environment is thus an ideal approach. Experimentation with new policies may be easily implemented at low cost and at little risk.

The principal objective of this section is to develop a model that is sufficiently detailed mathematically yet easily comprehensible to hospital managerial staff. The simulation model of the CCU was built using Visual Basic for Applications for Excel (VBA). The model seeks to simulate the bed occupancy of the CCU as well as considering various 'what if' scenarios. The main goals of 'what if' scenarios were reducing variation of bed occupancy and increasing throughput of patients in the CCU.

The model was constructed to allow emergency and elective patients to arrive at the CCU. Each patient type had a different arrival process and different length of stay profile. Detailed information about admission patterns and duration of stay profiles used for this model is described in Section 2.3.1 and 2.3.3.

In order to replicate the real-life situation in the CCU, the simulation model contains a total of 29 (24 funded and 5 unfunded) CCU beds. As shown in Section 2.3.4 on some occasions there were more than 29 beds occupied when patients were most likely have been admitted into so called 'virtual beds'. The simulation model does not allow any elective admissions when all beds are full and it allows fewer emergency patients to be admitted.

The simulation model generates the arrival of emergency and elective patients at the CCU, with each type of patient having a separate arrival process, as described in Section 2.3.1. To generate the number of arrivals of emergency and elective patients per day a Weighted Poisson distribution, with probability distribution function (PDF):

$$P(X = x) = \omega \frac{e^{-\alpha_1} {\alpha_1}^x}{x!} + (1 - \omega) \frac{e^{-\alpha_2} {\alpha_2}^x}{x!} \qquad x = 0, 1, 2, ...$$

is used, where ω is probability of given distribution and must be a number between 0 and 1; α_1, α_2 are the distribution parameters and both must be positive.

In the model, if a patient arrives at a time where there are unoccupied beds, they are admitted to the CCU. If an arriving elective patient finds that all beds are occupied, their surgery is cancelled and they are lost to the system (queueing is not allowed). If an arriving emergency patient finds that all beds are occupied they are admitted to a different ward, for example Cardiac Care Unit, or are held in the Recovery room following surgery.

Consider now the method by which the number of elective arrivals is generated in the model. As shown in Section 2.3.1, the number of arrivals is dependent on the day of the week. Additional analysis is necessary to obtain the appropriate distribution of the number of elective arrivals depending on the day of the week. It is concluded that a Weighted Poisson Distribution is appropriate to model the number of elective arrivals on each day of the week. The parameter values are found using *Solver* by minimising squares of deviation between data and model, and are detailed in Table 3.1.

The ω parameter value for Wednesday, Saturday and Sunday are either very close to 0 or to 1 suggesting that for these three days a simple Poisson Distribution would be appropriate. However, for consistency reasons it is decided to sample the number of admissions from the Weighted Poisson

Table 3.1: Parameter values for elective admissions depending on the day of the week

Day of the week	**Parameter values**
Monday	$\omega = 0.0869, \alpha_1 = 2.4948, \alpha_2 = 2.6540$
Tuesday	$\omega = 0.5156, \alpha_1 = 1.8343, \alpha_2 = 1.0505$
Wednesday	$\omega = 0.001, \alpha_1 = 1.7057, \alpha_2 = 1.7057$
Thursday	$\omega = 0.0393, \alpha_1 = 0.001, \alpha_2 = 2.0442$
Friday	$\omega = 0.9379, \alpha_1 = 1.5988, \alpha_2 = 3.2566$
Saturday	$\omega = 0.001, \alpha_1 = 0.4572, \alpha_2 = 0.4549$
Sunday	$\omega = 0.999, \alpha_1 = 0.2839, \alpha_2 = 2.0071$

distribution.

Sampling from a Weighted Poisson distribution is not direct since it is a weighted sum of two Poisson distributions with different parameters. The simulation model samples from the Poisson distribution with parameter α_1 with probability ω and from the Poisson distribution with parameter α_2 with probability $(1 - \omega)$. The simulation model generates a random number $\in (0, 1]$ from the Uniform distribution; then, if that random number is less than or equal to ω sampling from the Poisson distribution with parameter α_1 is undertaken; otherwise, sampling from the Poisson distribution with parameter α_2 is employed. Then, the next random number is generated to decide how many arrivals should occur, if any. Using the theoretical cumulative distribution function for the number of arrivals on each day of the week the number of admissions can be found. Depending on the value of the random number and on the day of week, the number of elective admissions varies from 0 up to 8 per day. Note that there is a restriction regarding admission of elective patients. For example, if the bed occupancy on the previous day was 24 and the simulated number of emergency admissions on the current day was four and the simulated number of elective admissions was three, only one elective admission would be allowed, so that the bed occupancy does not exceed 29.

Consider now the emergency arrivals. An analysis of these arrivals, detailed in Section 2.3.1, did not highlight any daily trends, thus there is no need to find the different parameters for each day of the week. The values of parameters that describe the Weighted Poisson distribution are obtained to be as follows: $\omega = 0.333$, $\alpha_1 = 0.5918$, $\alpha_2 = 2.6540$. If there are 29 beds occupied less emergency admissions will be permitted. Additional analysis was necessary to obtain an appropriate distribution of the number of emergency arrivals when the bed occupancy on the previous day exceeded 29. A separate distribution was fitted and it was concluded that the Poisson Distribution was appropriate to model the number of those arrivals; the PDF is as follows:

$$P(X = x) = \frac{e^{-\alpha}\alpha^x}{x!} \qquad x = 0, 1, 2, ...$$

The required α parameter that is found using *Solver* is $\alpha = 2.6624$. To obtain the number of emergency admissions per day the same procedure was used as explained above for elective patient case. The number of emergency admissions varies from 0 to 11 per day, and lies between 0 and 6 if the previous day bed occupancy was 29 or more.

Once the number of arrivals is calculated for each day it is necessary to simulate the length of stay for each patient admitted. The duration of stay that each patient is expected to spend receiving treatment in the CCU is highly dependent on the patient type, as described in Section 2.3.3. The distributions of the length of stay for emergency and elective patients are considered in detail in Section 2.3.3.1 and 2.3.3.2. The Weighted Negative Exponential distributions that were found to provide the best fit to the actual data are utilised in the simulation model. The PDF of the random variable x is:

$$f(x) = \omega\beta_1 e^{-\beta_1}x + (1-\omega)\beta_2 e^{-\beta_2 x}, \qquad x > 0$$

The parameters that are used are displayed in Table 3.2.

Table 3.2: Parameter values for the length of stay distribution

Patient type	**Parameter value**
Emergency	$\omega = 0.2993, \beta_1 = 0.1656$ (days), $\beta_2 = 0.9811$ (days)
Elective	$\omega = 0.2566, \beta_1 = 0.0636$ (days), $\beta_2 = 0.3039$ (days)

The simulation model samples from the Negative Exponential distribution with parameter β_1 with probability ω and from the Negative Exponential distribution with parameter β_2 with probability $(1-\omega)$. The simulation model samples a random number, then if that random number is less than or equal to ω sampling from the Negative Exponential distribution with parameter β_1 is undertaken and the length of stay (LoS) is rounded to the nearest integer, where LoS was generated using the inverse distribution method, as follows

$$LoS = \left(-\frac{1}{\beta_1}\right)\log(u)$$

where u is a random number in the range (0,1]. If the random number generated is zero then the simulation model re-samples it, since $\log(0)$ is not allowed.

If the generated random number is greater than ω then sampling from the Negative Exponential distribution with parameter β_2 is undertaken and the length of stay is LoS rounded to the next nearest integer, where

$$LoS = \left(-\frac{1}{\beta_2}\right)\log(u)$$

Once the arrival day and length of stay is generated for each patient, the discharge date can be easily obtained.

And, finally, the bed occupancy on each day is calculated by the following formula:

bed occupancy at midnight today = bed occupancy at midnight yesterday + number of emergency admissions today + number of elective admissions today - number of discharges today.

3.2.1 Validation of the Simulation Model

Prior to obtaining results from the simulation model and drawing conclusions based on these results, assurance that the model provides good representation of the real-life situation is required.

The simulation model can not start with an empty system (i.e. no patients occupying any of the beds). This is not representative of the real-life situation; the CCU is not emptied at the end of each year. Therefore, the number of beds occupied at the beginning of each run would be disproportionately low. In order to prevent these early readings influencing the overall mean outputs, it was decided to start the system with some beds occupied; the integer random number between 0 and 29 was generated by the simulation model. As stated before, on average $\frac{2}{3}$ of all patients are emergency, so the number of beds that are occupied by emergency patients is the integer part of the product $\frac{2}{3}\times$(random number between 0 and 29). The number of beds that are occupied by elective patients is simply that generated by the above random number, and then subtracting the number of emergency patients. In a similar way, as explained previously, the length of stay is generated for both patient types.

Running the simulation model for an insufficient number of days may produce unreliable results. This is because each run of the simulation model, using different random numbers, produces varying results. To allow the model to enter steady state conditions, a warm up period equivalent to 3 months is included. Obviously, the greater the number of runs in a trial, the more accurate the results. It is decided to run the model for 1,000,000 days to ensure stability in the results. It is a sufficiently long period given the time to run.

As described previously, the model was constructed so that two types of patients arrive at the CCU and are served according to a statistical distribution based on patient type. The arrivals and LoS of each patients' type will be considered independently.

3.2.1.1 Emergency Number of Arrivals

The distribution of the number of emergency admissions was analysed from the simulated numbers of arrivals on each day and compared with the data. These are presented graphically in Figure 3.1.

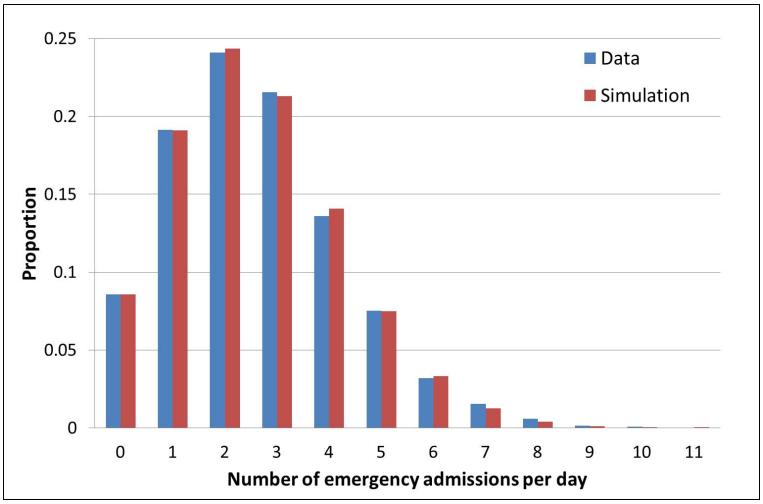


Figure 3.1: Distribution of the daily emergency admissions

Clearly, the simulation model gives a very good fit to the distribution of emergency arrivals. From the data, the mean number of emergency admissions per day is 2.61 which compares favourably with the simulation mean of 2.59. It can be concluded that the Weighted Poisson Distribution does provide a very good fit to the data.

3.2.1.2 Elective Number of Arrivals

The distribution of the simulated number of elective admissions is compared with the data. These are illustrated in Figure 3.2.

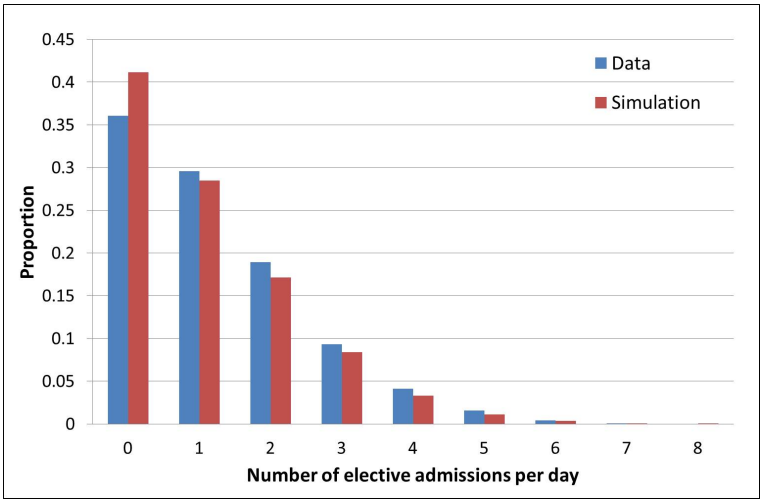


Figure 3.2: Distribution of the daily elective admissions

The simulation model slightly overestimates the number of days when there were no elective admissions. From the data, the mean number of elective admissions is 1.22 and 1.09 from simulation.

However, it is concluded that the overall number of elective admissions distribution is modelled sufficiently using the Weighted Poisson distribution.

Overall it is concluded that the simulation model provides satisfactory arrival numbers for both emergency and elective patients. The length of stay distribution for emergency and elective patients will also be considered separately.

3.2.1.3 Emergency Length of Stay

The distribution of length of stay of emergency patients was produced from the simulated lengths of stay and compared with the actual data. These are presented graphically in Figure 3.3.

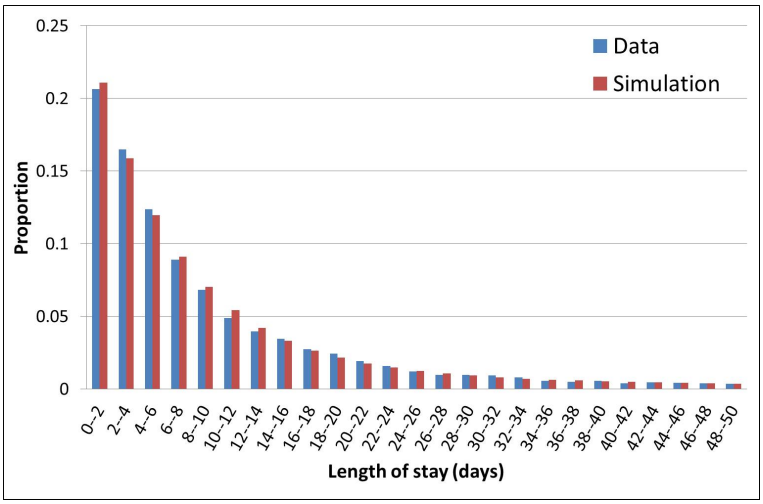


Figure 3.3: Distribution of emergency length of stay

Visually, the simulation model gives a very good fit to the distribution of emergency length of stay. From the data, the average length of stay is 7.00 days which compares favourably with the simulation mean of 6.93. It can be confirmed that the Weighted Negative Exponential distribution does fit the data very well.

3.2.1.4 Elective Length of Stay

The distribution of simulated elective length of stay is compared with the data. These are illustrated in Figure 3.4. Clearly, the simulation model provides a very good fit to the distribution of elective length of stay. From the data, the average length of stay is 3.08 days which compares favourably with the simulation mean of 2.93. It can be concluded that the Weighted Negative Exponential distribution does fit the data very well.

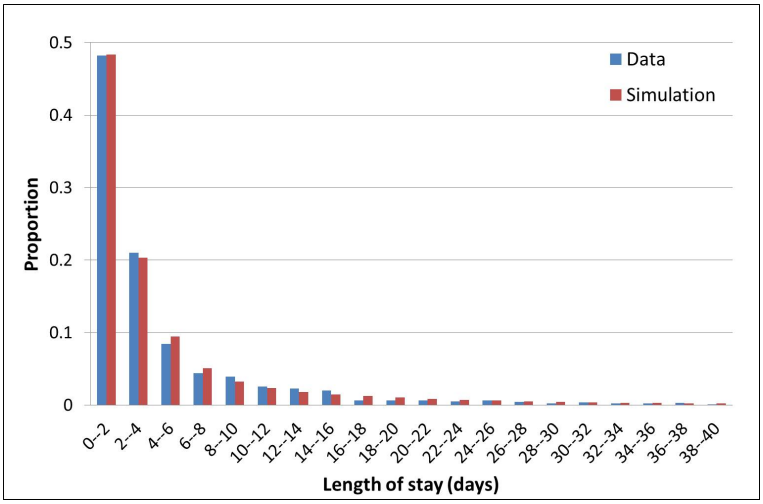


Figure 3.4: Distribution of elective length of stay

3.2.2 Results of the Simulation Model

Having developed a simulation model of the CCU based on actual data and having validated some key outcomes of the model, it can now be utilised as an operational research tool.

The final result consists of comparing the bed occupancy profile from the data with the output from the model. The model is run with 29 beds available. This resulted in 74% bed occupancy utilisation rate compared with observed rate of 75%. Recall that on 1.32% of occasions there were more than 29 beds occupied. In the simulation model it appeared that on 2.27% of occasions bed occupancy was higher than 29, and in fact there were as many as 36 patients at any one time. The main reason for this is the same as in the data, patients are allowed to queue on trolleys if the hospital staff know there will be a bed available in the CCU shortly.

The measures that are examined are: mean and standard deviation of bed occupancy. The mean and the standard deviation of bed occupancy, according to the data, was 21.85 and 3.65 respectively, compared with 21.51 and 4.28 respectively, according to the simulation. The results of the simulation model are compared with the data and are presented in Figure 3.5.

The bed occupancy distribution displayed in Figure 3.5 is comparable with the actual bed occupancy profile, and hence it is concluded that the simulation model provides a reasonably accurate representation of the real-life CCU activities. However, the simulation model overestimates the low bed occupancy (between 0 and 18) and high bed occupancy (between 27 and 35) and underestimates mid-valued bed occupancy, which is the reason for a slightly higher standard deviation comparing with the actual data. In order to investigate that variation in bed occupancy further,

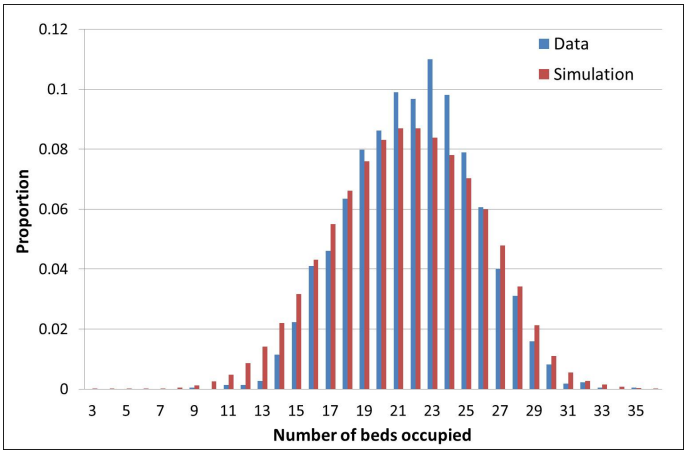


Figure 3.5: Distribution of bed occupancy

several 'what if' scenarios will be considered.

3.2.3 'What if' Scenario #1

The previous section detailed some initial results of the simulation model; however, no major alternations were made to the model; arrival rates and duration of stay remained unchanged.

This section will examine the effect of implementing some new policies regarding cancellation of elective surgeries. As has been previously mentioned, the CCU has an insufficient number of beds to accommodate demand on all occasions, and sometimes elective surgeries may require cancellation. Recall that the main method of control over the rate of admissions to the CCU is by means of changing the admission rates of elective patients. A principle of the design of the simulation model is that it does not allow any elective admissions when bed occupancy reaches high levels. A point at which elective arrivals are beginning to be cancelled will be called the 'cut-off' point. Since the number of funded beds in the CCU is 24, when bed occupancy is 24 or higher no elective admissions are allowed. The proposed rule for the cancellation of elective procedures during busy periods is incorporated into the model. If on any one day after admitting priority emergency patients, the number of occupied beds was greater or equal to 24, the number of elective admissions was set to zero. It means that all planned elective surgeries for that day are cancelled and those patients are lost to the system, since in this model no queueing is allowed.

The effect of that 'what if' scenario is highly influential. The measures that were again examined are: mean and standard deviation of bed occupancy. The mean number and the standard deviation of bed occupancy, according to the simulation, is 20.24 and 3.73 respectively. Unsurprisingly, the average bed occupancy is now lower than in the data, since the model does not allow as many elective admissions as previously. This is shown in Figure 3.6.

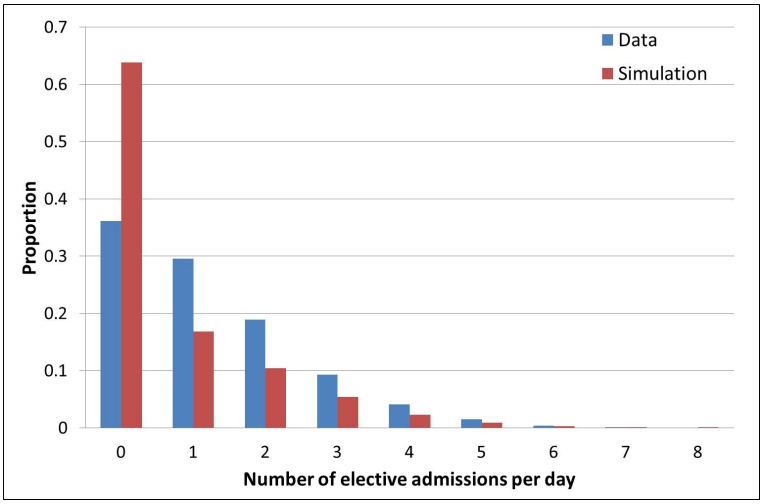


Figure 3.6: Distribution of number of elective admissions with cut-off at 24

Clearly, the percentage of days when there were no elective admissions is very high (64%) and the effect on bed occupancy is presented in Figure 3.7.

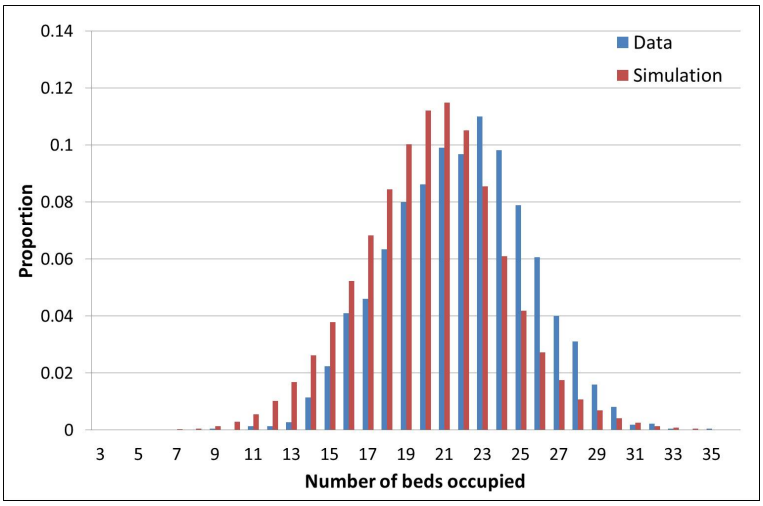


Figure 3.7: Distribution of bed occupancy with cut-off at 24

High bed occupancies are no longer overestimated by simulation, but clearly the proportion of low bed occupancy is overestimated. This would be undesirable since each bed is a very expensive and limited resource, and there are many people who have to wait to be admitted to the CCU for post-operative treatment. Also, if for example, there are 20 nurses employed per shift and only ten beds are occupied, there is potential for a large wastage in nursing cost. To rectify this problem a second 'what if' scenario is considered.

The implication of changes in the mode of operation of the CCU is now considered. Recall also that the main method of control over the rate of admissions to the CCU is by means of changing

the admission rates of elective patients. Recall that two of the principal aims of this study are to increase the throughput of patients, and to reduce the variation in bed occupancy levels on a day-to-day basis, so that more stability may be observed in the numbers of nurses needed per shift.

3.2.4 'What if' Scenario #2

The second 'what if' scenario investigates the effect of increasing the number of elective admissions by up to 4 per day whenever there are less than 24 beds occupied. That is, whenever there appears to be sufficient spare bed capacity, then allowing extra (up to four) elective patients to be admitted is suggested. For example, if there are currently 22 beds occupied, only two extra elective patients would be allowed, since the cut-off at 24 beds. However, if there are 20 or less beds in use four extra admissions are allowed. The question arose: is it realistic to expect patients facing elective surgery to have that surgery brought forward at short notice (typically three days)? Discussions with clinical staff indicate that this indeed may be possible. For example, transplant patients are aware that they may be called for surgery at a very short notice if a donated organ becomes available. Likewise, it may be possible to set up a pool of patients waiting for more general surgery, with agreement reached beforehand with patients that their surgery may be performed sooner if they agree to join the pool for call-up at three days or so notice.

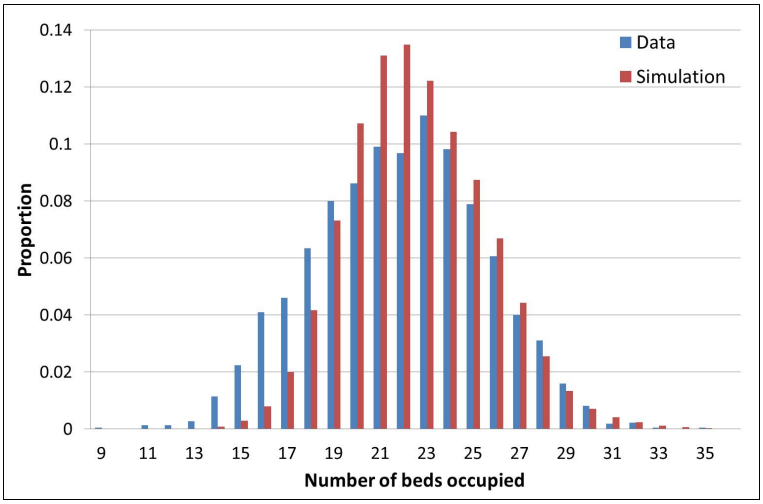


Figure 3.8: Distribution of bed occupancy with cut-off at 24 and extra elective admissions at non-busy times

The results of the simulation model are compared with the data and are illustrated in Figure 3.8. Visibly, very low and very high bed occupancy levels are no longer overestimated and mid-valued bed occupancy levels are higher than in the actual data. The mean bed occupancy increases to 22.6, a 3.5% increase and the standard deviation is reduced to 3.00, a 17.8% decrease. The variation of bed occupancy levels on a day-to-day basis was reduced significantly, which makes the system more stable and decisions regarding the number of nurses needed per shift are easier to make. The

throughput is increased from 1406 to 1473 patients per year, a 4.8% increase. Thus, a relatively minor increase in admissions of elective patients at non-busy times shows a improvement in variability and throughput at no extra cost to the hospital.

3.2.5 Conclusions

The principal objective of this section was to develop a model that is sufficiently detailed mathematically and easily comprehensible to hospital managerial staff. It is believed that this objective has been achieved. Detailed analyses of arrival and length of stay profiles, presented in Section 2.3, provided results of the simulation model of the CCU that correspond relatively well with the actual CCU data.

Consider now a theoretical approach, whose objectives are the same as of the simulation model.

3.3 Analytical Model

3.3.1 Introduction

From the previous discussion, it is suggested that any mathematical model must cater for the two categories of patients, both with regard to their admission rate and length of stay. Further, the hospital states that it never permits a queue to occur for admission to the CCU. At first sight, this appears to be an unlikely scenario. The explanation is that to avoid a queue forming the hospital would either temporarily cancel elective admissions, make attempts to divert a potential admission elsewhere, or create a 'virtual bed'. This can be described as a trolley bed, possibly located outside the CCU environment. With these matters in mind, a simple service model with no queueing allowed is initially proposed.

This section will commence by considering a so called $M_2/M_2/c/c/FIFO$ queue, taken to be a system with random arrivals from two different streams and corresponding negative exponential service times (length of stay) depending on patients' type, c service channels (beds), capacity of the system c and first-in-first-out queueing discipline.

3.3.2 The Queueing Model $M_2/M_2/c/c/FIFO$

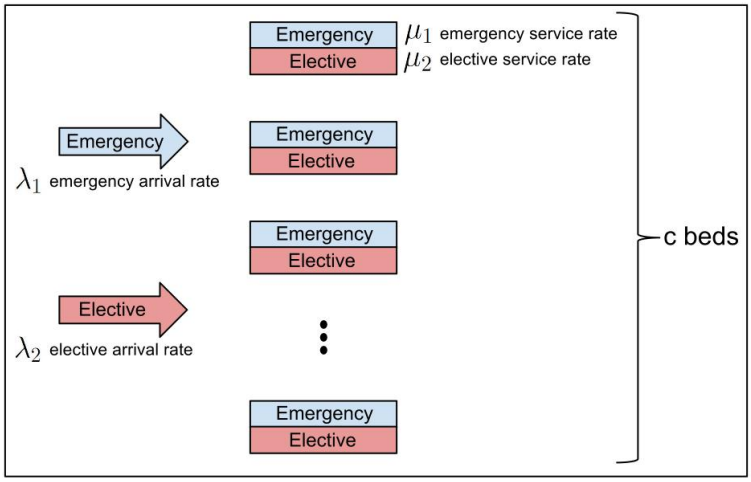


Figure 3.9: Schematic diagram illustrating the principal features of the queueing model

Figure 3.9 illustrates the system. Emergency patients arrive at random at mean rate λ_1, and elective patients arrive at random at mean rate λ_2. The mean length of stay of an emergency patient is denoted by $\frac{1}{\mu_1}$, while that of elective patients is denoted by $\frac{1}{\mu_2}$. There are c service channels (beds) available. Later, the application of this queueing model will be shown when there are 29 service channels available. No queues are allowed to form. The objective at this stage therefore is to determine how daily bed occupancy varies, and importantly to investigate how that occupancy is distributed amongst emergency and elective patients.

Let $P_{i,j}(t)$ denote the probability that i emergency and j elective patients are present in the system at time t. The various states of the system are denoted by the double suffix (i, j) for $0 \leq i + j \leq c$ and $i, j \in \mathbb{N}$, where i corresponds to the number of emergency patients and j corresponds to the number of elective patients. If there were no beds in use the only possible state of the system would be: $(0, 0)$, so zero emergency and zero elective patients. If there was one bed in use, the two possible states of the system would be: $(1, 0)$ or $(0, 1)$, so one emergency and no electives or no emergencies and one elective patient. If there were two beds in use, the three possible states are: $(2, 0), (1, 1)$ or $(0, 2)$. If there were three beds in use, the four possible states are: $(3, 0), (2, 1), (1, 2)$ or $(0, 3)$. Therefore if the system had only 3 beds available, the number of possible states of the system would be: $1 + 2 + 3 + 4 = 10$. As there are c beds available, this leads to $\sum_{r=0}^{c} (r + 1) = \frac{(c+1)(c+2)}{2}$ possible states of the system. For example for 29 beds there are 465 possible states of the system. The overall bed occupancy probability, $P_n(t)$, at time t is given by:

$$P_n(t) = \sum_{i=0}^{n} P_{i,n-i}(t) \qquad \text{for } n = 0, 1, 2, ..., c$$

The differential-difference equations to describe the system for c service channels (number of beds) are set-up. The $\sum_{r=0}^{c}(r+1)$ equations can be categorised into seven types of the differential-difference equations describing the system. They depend on the values of i and j and are as follows:

(1) For $i = j = 0$:

$$\begin{aligned}P_{0,0}(t+\delta t) =& P_{0,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t] \\ &+ P_{1,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]\mu_1\delta t \\ &+ P_{0,1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]\mu_2\delta t + o(\delta t)\end{aligned}$$

(2) For $1 \leq i \leq (c-1)$ and $j = 0$:

$$\begin{aligned}P_{i,0}(t+\delta t) =& P_{i,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-i\mu_1\delta t] \\ &+ P_{i-1,0}(t)[\lambda_1\delta t][1-\lambda_2\delta t][1-(i-1)\mu_1\delta t] \\ &+ P_{i+1,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](i+1)\mu_1\delta t \\ &+ P_{i,1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-i\mu_1\delta t]\mu_2\delta t + o(\delta t)\end{aligned}$$

(3) For $i = 0$ and $1 \leq j \leq (c-1)$:

$$\begin{aligned}P_{0,j}(t+\delta t) =& P_{0,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-j\mu_2\delta t] \\ &+ P_{0,j-1}(t)[1-\lambda_1\delta t][\lambda_2\delta t][1-(j-1)\mu_2\delta t] \\ &+ P_{0,j+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](j+1)\mu_2\delta t \\ &+ P_{1,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][\mu_1\delta t][1-j\mu_2\delta t] + o(\delta t)\end{aligned}$$

(4) For $i + j \leq (c-1)$ and $i, j \neq 0$:

$$\begin{aligned}P_{i,j}(t+\delta t) =& P_{i,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-i\mu_1\delta t][1-j\mu_2\delta t] \\ &+ P_{i-1,j}(t)[\lambda_1\delta t][1-\lambda_2\delta t][1-(i-1)\mu_1\delta t][1-j\mu_2\delta t] \\ &+ P_{i,j-1}(t)[1-\lambda_1\delta t][\lambda_2\delta t][1-i\mu_1\delta t][1-(j-1)\mu_2\delta t] \\ &+ P_{i+1,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][(i+1)\mu_1\delta t][1-j\mu_2\delta t] \\ &+ P_{i,j+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-i\mu_1\delta t][(j+1)\mu_2\delta t] + o(\delta t)\end{aligned}$$

(5) For $i = 0$ and $j = c$:

$$\begin{aligned}P_{0,c}(t+\delta t) =& P_{0,c}(t)[1-c\mu_2\delta t] \\ &+ P_{0,c-1}(t)[1-\lambda_1\delta t][\lambda_2\delta t][1-(c-1)\mu_2\delta t] + o(\delta t)\end{aligned}$$

(6) For $i = c$ and $j = 0$:

$$\begin{aligned} P_{c,0}(t+\delta t) =& P_{c,0}(t)[1 - c\mu_1\delta t] \\ &+ P_{c-1,0}(t)[\lambda_1\delta t][1 - \lambda_2\delta t][1 - (c-1)\mu_1\delta t] + o(\delta t) \end{aligned}$$

(7) For $i + j = c$ and $i, j \neq 0$:

$$\begin{aligned} P_{i,j}(t+\delta t) =& P_{i,j}(t)[1 - i\mu_1\delta t][1 - j\mu_2\delta t] \\ &+ P_{i-1,j}(t)[\lambda_1\delta t][1 - \lambda_2\delta t][1 - (i-1)\mu_1\delta t][1 - j\mu_2\delta t] \\ &+ P_{i,j-1}(t)[1 - \lambda_1\delta t][\lambda_2\delta t][1 - i\mu_1\delta t][1 - (j-1)\mu_2\delta t] + o(\delta t) \end{aligned} \tag{3.1}$$

At this point, only the steady-state solutions of these equations are considered; the time-dependent aspect will be considered later in Section 4.2.

The steady-state equations may be written in the form shown below:

(1) For $i = j = 0$:

$$(\lambda_1 + \lambda_2)P_{0,0} = \mu_1 P_{1,0} + \mu_2 P_{0,1}$$

(2) For $1 \leq i \leq (c-1)$ and $j = 0$:

$$(\lambda_1 + \lambda_2 + i\mu_1)P_{i,0} = \lambda_1 P_{i-1,0} + (i+1)\mu_1 P_{i+1,0} + \mu_2 P_{i,1}$$

(3) For $i = 0$ and $1 \leq j \leq (c-1)$:

$$(\lambda_1 + \lambda_2 + j\mu_2)P_{0,j} = \lambda_2 P_{0,j-1} + (j+1)\mu_2 P_{0,j+1} + \mu_1 P_{1,j}$$

(4) For $i + j \leq (c-1)$ and $i, j \neq 0$:

$$(\lambda_1 + \lambda_2 + i\mu_1 + j\mu_2)P_{i,j} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + (i+1)\mu_1 P_{i+1,j} + (j+1)\mu_2 P_{i,j+1}$$

(5) For $i = 0$ and $j = c$:

$$c\mu_2 P_{0,c} = \lambda_2 P_{0,c-1}$$

(6) For $i = c$ and $j = 0$:

$$c\mu_1 P_{c,0} = \lambda_1 P_{c-1,0}$$

(7) For $i + j = c$ and $i, j \neq 0$:

$$(i\mu_1 + j\mu_2)P_{i,j} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} \tag{3.2}$$

Theorem 3.3.1. *The steady-state probability that there are i emergency and j elective patients present with different arrival rates (λ_1 and λ_2) and different service rates (μ_1 and μ_2) in the system is given by:*

$$P_{i,j} = \frac{1}{i!j!}{\theta_1}^i{\theta_2}^j P_{0,0} \qquad (i = 0, 1, ..., c; j = 0, 1, 2, ..., c - i)$$

where

$$\theta_k = \frac{\lambda_k}{\mu_k}, \qquad k = 1, 2$$

$$P_{0,0} = \frac{1}{\sum_{i=0}^{c}\sum_{j=0}^{c-i}\frac{1}{i!j!}{\theta_1}^i{\theta_2}^j}$$

Proof.
The proof of this result is relatively straightforward, using an inductive approach. The required algebraic manipulation for each of the steady-state Equations 3.2 is given below.

(1) For $i = j = 0$:

$$\begin{aligned}
P_{0,0} =& \frac{\mu_1 P_{1,0} + \mu_2 P_{0,1}}{\lambda_1 + \lambda_2} && \text{using steady-state equation (1)}\\
=& \frac{\mu_1\theta_1 P_{0,0} + \mu_2\theta_2 P_{0,0}}{\lambda_1 + \lambda_2} && \text{using the expression of Theorem 3.3.1}\\
=& \frac{\mu_1\frac{\lambda_1}{\mu_1} + \mu_2\frac{\lambda_2}{\mu_2}}{\lambda_1 + \lambda_2}P_{0,0}\\
=& \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2}P_{0,0}\\
=& P_{0,0} && \square
\end{aligned}$$

(2) For $1 \leq i \leq (c-1)$ and $j = 0$:

$$\begin{aligned}
P_{i,0} =& \frac{\lambda_1 P_{i-1,0} + (i+1)\mu_1 P_{i+1,0} + \mu_2 P_{i,1}}{\lambda_1 + \lambda_2 + i\mu_1} && \text{using steady-state equation (2)} \\
=& \frac{\lambda_1 \frac{1}{(i-1)!}{\theta_1}^{i-1} P_{0,0} + (i+1)\mu_1 \frac{1}{(i+1)!}{\theta_1}^{i+1} P_{0,0} + \mu_2 \frac{1}{i!}{\theta_1}^{i}\theta_2 P_{0,0}}{\lambda_1 + \lambda_2 + i\mu_1} \\
=& \frac{\frac{1}{i!}{\theta_1}^{i}(\frac{i\lambda_1}{\theta_1} + \mu_1\theta_1 + \mu_2\theta_2) P_{0,0}}{\lambda_1 + \lambda_2 + i\mu_1} \\
=& \frac{\frac{1}{i!}{\theta_1}^{i}(i\mu_1 + \lambda_1 + \lambda_2) P_{0,0}}{\lambda_1 + \lambda_2 + i\mu_1} \\
=& \frac{1}{i!}{\theta_1}^{i} P_{0,0} && \square
\end{aligned}$$

(3) For $i = 0$ and $1 \leq j \leq (c-1)$:

$$\begin{aligned}
P_{0,j} =& \frac{\lambda_2 P_{0,j-1} + (j+1)\mu_2 P_{0,j+1} + \mu_1 P_{1,j}}{\lambda_1 + \lambda_2 + j\mu_2} && \text{using steady-state equation (3)} \\
=& \frac{\lambda_2 \frac{1}{(j-1)!}{\theta_2}^{j-1} P_{0,0} + (j+1)\mu_2 \frac{1}{(j+1)!}{\theta_2}^{j+1} P_{0,0} + \mu_1 \frac{1}{j!}\theta_1{\theta_2}^{j} P_{0,0}}{\lambda_1 + \lambda_2 + j\mu_2} \\
=& \frac{\frac{1}{j!}{\theta_2}^{j}(\frac{j\lambda_2}{\theta_2} + \mu_2\theta_2 + \mu_1\theta_1) P_{0,0}}{\lambda_1 + \lambda_2 + j\mu_2} \\
=& \frac{\frac{1}{j!}{\theta_2}^{j}(j\mu_2 + \lambda_2 + \lambda_1) P_{0,0}}{\lambda_1 + \lambda_2 + j\mu_2} \\
=& \frac{1}{j!}{\theta_2}^{j} P_{0,0} && \square
\end{aligned}$$

(4) For $i + j \leq (c-1)$ and $i, j \neq 0$:

$$\begin{aligned}
P_{i,j} =& \frac{\lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + (i+1)\mu_1 P_{i+1,j} + (j+1)\mu_2 P_{i,j+1}}{\lambda_1 + \lambda_2 + i\mu_1 + j\mu_2} && \text{using steady-state equation (4)} \\
=& \frac{\frac{\lambda_1}{(i-1)!j!}{\theta_1}^{i-1}{\theta_2}^{j} + \frac{\lambda_2}{i!(j-1)!}{\theta_1}^{i}{\theta_2}^{j-1} + \frac{(i+1)\mu_1}{(i+1)!j!}{\theta_1}^{i+1}{\theta_2}^{j} + \frac{(j+1)\mu_2}{i!(j+1)!}{\theta_1}^{i}{\theta_2}^{j+1}}{\lambda_1 + \lambda_2 + i\mu_1 + j\mu_2} P_{0,0} \\
=& \frac{\frac{1}{i!j!}{\theta_1}^{i}{\theta_2}^{j}\left(\frac{i\lambda_1}{\theta_1} + \frac{j\lambda_2}{\theta_2} + \mu_1\theta_1 + \mu_2\theta_2\right)}{\lambda_1 + \lambda_2 + i\mu_1 + j\mu_2} P_{0,0} \\
=& \frac{\frac{1}{i!j!}{\theta_1}^{i}{\theta_2}^{j}(i\mu_1 + j\mu_2 + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + i\mu_1 + j\mu_2} P_{0,0} \\
=& \frac{1}{i!j!}{\theta_1}^{i}{\theta_2}^{j} P_{0,0} && \square
\end{aligned}$$

(5) For $i = 0$ and $j = c$:

$$\begin{aligned} P_{0,c} =& \frac{\lambda_2 P_{0,c-1}}{c\mu_2} && \text{using steady-state equation (5)} \\ =& \frac{\lambda_2 \frac{1}{(c-1)!} {\theta_2}^{c-1} P_{0,0}}{c\mu_2} \\ =& \frac{1}{c!} {\theta_2}^{c} P_{0,0} && \square \end{aligned}$$

(6) For $i = c$ and $j = 0$:

$$\begin{aligned} P_{c,0} =& \frac{\lambda_1 P_{c-1,0}}{c\mu_1} && \text{using steady-state equation (6)} \\ =& \frac{\lambda_1 \frac{1}{(c-1)!} {\theta_1}^{c-1} P_{0,0}}{c\mu_1} \\ =& \frac{1}{c!} {\theta_1}^{c} P_{0,0} && \square \end{aligned}$$

(7) For $i + j = c$ and $i, j \neq 0$:

$$\begin{aligned} P_{i,j} =& \frac{\lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1}}{i\mu_1 + j\mu_2} && \text{using steady-state equation (7)} \\ =& \frac{\lambda_1 \frac{1}{(i-1)!j!} {\theta_1}^{i-1} {\theta_2}^{j} P_{0,0} + \lambda_2 \frac{1}{i!(j-1)!} {\theta_1}^{i} {\theta_2}^{j-1} P_{0,0}}{i\mu_1 + j\mu_2} \\ =& \frac{\frac{1}{i!j!} {\theta_1}^{i} {\theta_2}^{j} \left(\frac{i\lambda_1}{\theta_1} + \frac{j\lambda_2}{\theta_2} \right)}{i\mu_1 + j\mu_2} P_{0,0} \\ =& \frac{\frac{1}{i!j!} {\theta_1}^{i} {\theta_2}^{j} \left(i\mu_1 + j\mu_2 \right)}{i\mu_1 + j\mu_2} P_{0,0} \\ =& \frac{1}{i!j!} {\theta_1}^{i} {\theta_2}^{j} P_{0,0} && \square \quad (3.3) \end{aligned}$$

This shows that the expression of Theorem 3.3.1 is a solution to the steady-state equations. Uniqueness of this solution follows from the fact that the solution to these equations correspond to the stationary distribution of the continuous Markov chain which can be expressed in matrix form as:

$$\pi Q = 0 \text{ with } \pi e = 1$$

for the corresponding stochastic matrix Q. A continuous-time Markov chain can be discretized since $\pi(Q\Delta t + I) = \pi$ and hence π is the eigenvector corresponding to the eigenvalue $\lambda = 1$ for the transition probability matrix of a finite discrete-time Markov chain. Therefore by the property of stochastic matrix there exist unique π such that $\sum_{i=1}^{n} \pi_i = 1$ [152].

Now the results obtained for $P_{i,j}$ (probability of having i emergency patients and j elective patients) are compared with the data in Figure 3.10.

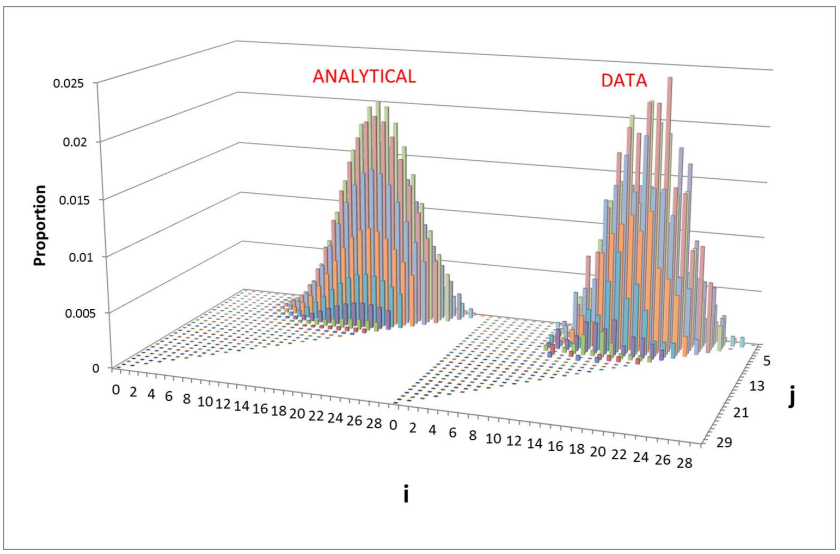


Figure 3.10: Comparison of analytical results with the data for $P_{i,j}$

It might be visually difficult to compare both graphs, therefore the difference between the data and analytical results is obtained, and Figure 3.11 presents the difference.

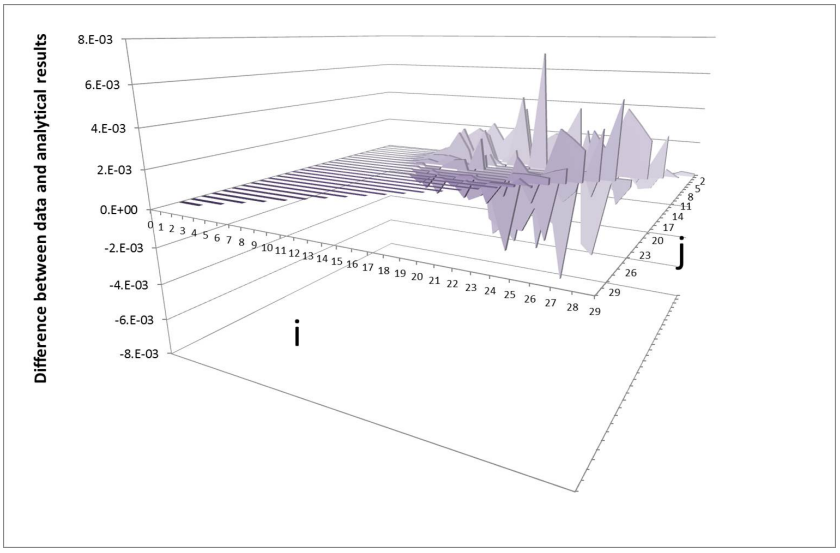


Figure 3.11: Difference between analytical results and the data for $P_{i,j}$

Figure 3.10 shows close agreement when comparing the results from Theorem 3.3.1 and the data for probability of having i emergency patients and j elective patients present in the system. Figure 3.11 confirms it; the maximum absolute value of the difference is 0.00704.

An application of Theorem 3.3.1 is offered in Section 3.3.4.

The probability of the number of beds occupied, P_n, (irrespective of whether the occupants are emergencies or electives) is obtained below:

$$\begin{aligned} P_n &= \sum_{i=0}^{n} P_{i,n-i} \\ &= \sum_{i=0}^{n} \frac{1}{i!(n-i)!} {\theta_1}^i {\theta_2}^{n-i} P_0 \qquad \text{(using Theorem 3.3.1)} \\ &= \frac{1}{n!} P_0 \sum_{i=0}^{n} \binom{n}{i} {\theta_1}^i {\theta_2}^{n-i} \\ &= \frac{1}{n!} P_0 (\theta_1 + \theta_2)^n \\ &= \frac{1}{n!} \theta^n P_0 \end{aligned} \tag{3.4}$$

where:

$$\theta = \theta_1 + \theta_2, \qquad \theta_1 = \frac{\lambda_1}{\mu_1}, \qquad \theta_2 = \frac{\lambda_2}{\mu_2}$$

P_0 may then be evaluated in the usual way, using the fact that:

$$\sum_{n=0}^{c} P_n = 1$$

Therefore:

$$P_0 = \frac{1}{\sum_{n=0}^{c} \frac{1}{n!} \theta^n}$$

Now, the obtained results for P_n are compared with the data. A program written in Visual Basic calculates P_n for $n = 0, 1, 2, \ldots, 29$ using Formula 3.4. The parameters that are needed are: λ_1, λ_2, μ_1 and μ_2. The arrival and service rates that were described in Section 2.3 are now given in Table 3.3.

Table 3.3: Parameter values

Parameter	**Value**
λ_1	2.6081
λ_2	1.2231
μ_1	0.1406
μ_2	0.3260

The results are now compared in Figure 3.12, which shows close agreement in the overall bed occupancy levels when comparing the analytical results using Formula 3.4 with the actual data.

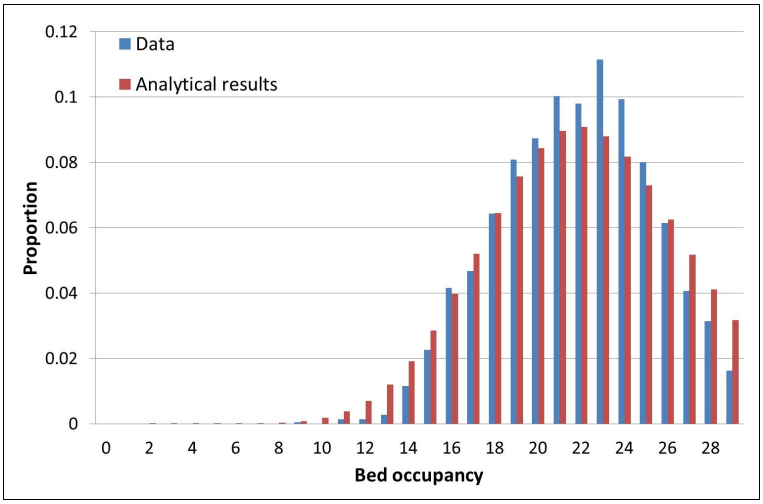


Figure 3.12: Comparison of analytical results with the actual data

Note that the maximum bed occupancy is 29. However, as mentioned before in Section 2.3.4 it can be as high as 35, thus the data needed standardising, so that the bed occupancy probabilities between 0 and 29 sum to one.

By standardizing the data, the mean and standard deviation of bed occupancy change. They are now 21.73 and 3.52 respectively, which compares favourably with the analytical results (mean of 21.59 and standard deviation of 4.04); however, the variation in bed occupancy is now slightly higher.

Note that Formula 3.4 is in fact Erlang's Loss Formula for the $M/M/c/c$ queue (Stewart, 2009 [152]). Further investigation is undertaken regarding the connection between P_n and $P_{i,j}$. In fact, the two independent admission streams (emergency and elective) may be combined into a single input. Also, the two service rates may also be combined into a single service rate.

As previously stated λ_1 and λ_2 are two different arrival rates. Given the parameters, the overall mean arrival rate is $\lambda = \lambda_1 + \lambda_2$. The mean service rate μ can be obtained by observing that the mean service time $\frac{1}{\mu_1}$ is experienced by a patient with probability $\frac{\lambda_1}{\lambda_1+\lambda_2}$, and $\frac{1}{\mu_2}$ by a patient with probability $\frac{\lambda_2}{\lambda_1+\lambda_2}$ and so the mean service time is:

$$\begin{aligned}
\frac{1}{\mu} &= \frac{1}{\mu_1}\frac{\lambda_1}{\lambda_1+\lambda_2} + \frac{1}{\mu_2}\frac{\lambda_2}{\lambda_1+\lambda_2} && \text{(as a weighted mean)} \\
&= \frac{\lambda_1\mu_2 + \lambda_2\mu_1}{\mu_1\mu_2(\lambda_1+\lambda_2)} \\
&= \frac{\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}}{\lambda_1+\lambda_2} \\
&= \frac{\theta_1+\theta_2}{\lambda} && \text{(letting } \lambda = \lambda_1+\lambda_2 \text{ and } \theta_k = \frac{\lambda_k}{\mu_k} \text{ for } k = 1,2) \qquad (3.5)
\end{aligned}$$

This gives the mean service rate:

$$\mu = \frac{\lambda}{\theta_1 + \theta_2}$$

and the steady state probabilities for the total number of customers in the system can be obtained using well known results for the $M/M/c/c$ queue with $\theta = \theta_1 + \theta_2$.

3.3.3 The queueing model $M_2/M/c/c+m/FIFO$

It has been seen from the data on bed occupancy that on some occasions (1.32%) the number of beds occupied exceeds the number of beds available in the CCU. The CCU manager explained that this could be due to a number of possibilities. For example, a patient could be treated in a general ward temporarily with a critical care nurse in attendance, or a patient having finished surgery is held in the Recovery room with an anaesthetist in attendance. For this reason, it is therefore worth looking at the following type of queueing system: $M/M/c/c+m/FIFO$, where m is small and $m \in \mathbb{N}$.

Patients arrive at random at mean rate λ, where $\lambda = \lambda_1 + \lambda_2$ and the overall mean LoS is denoted by $\frac{1}{\mu}$ and is given by Formula 3.5. There are c service channels and m available spaces in the queue. Let P_n denote the probability that n patients are present in the system. As there are $c + m$ beds available, this leads to $c + m + 1$ states of the system. They can be categorised into five types of the differential-difference equations describing the system with c service channels and additional m spaces in the queue. The formula for P_n is known and given in Stewart, 2009 [152] and is as follows:

$$P_n = \begin{cases} \frac{1}{n!}\theta^n P_0 & \text{if } 1 \leq n \leq c \\ \frac{1}{c^{n-c}c!}\theta^n P_0 & \text{if } c \leq n \leq c+m \end{cases}$$

and

$$P_0 = \frac{1}{\sum_{n=0}^{c-1} \frac{1}{n!}\theta^n + \sum_{n=c}^{c+m} \frac{1}{c^{n-c}c!}\theta^n}$$

where

$$\theta = \theta_1 + \theta_2 = \frac{\lambda_1}{\mu} + \frac{\lambda_2}{\mu}$$

A more complex system with two separate arrival streams of patients and combined service rate will be now considered. The system is similar to the system illustrated in Figure 3.9, but now patients are allowed to be admitted even if all beds are occupied. Emergency patients still arrive at random at mean rate λ_1, and elective patients still arrive at random at mean rate λ_2. The overall mean LoS is denoted by $\frac{1}{\mu}$ and is given by Formula 3.5. There are c service channels available with additional m spaces in the queue.

Let $P_{i,j}(t)$ denote the probability that i emergency and j elective patients are present in the system at time t. The various states of the system are denoted by the double suffix (i, j) for $0 \leq i+j \leq c+m$ and $i, j \in \mathbb{N}$. As there are $c + m$ beds available, this leads to $\sum_{r=0}^{c+m}(r + 1) = \frac{(c+m+1)(c+m+2)}{2}$ states of the system.

The $\sum_{r=0}^{c+m}(r + 1)$ equations can be categorised into 13 types of the differential-difference equations describing the system with c service channels and additional m spaces in the queue. They depend on the values of i and j and are as follows:

(1) For $i = j = 0$:

$$\begin{aligned}P_{0,0}(t+\delta t) =& P_{0,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t] \\ &+ P_{1,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]\mu\delta t \\ &+ P_{0,1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]\mu\delta t + o(\delta t)\end{aligned}$$

(2) For $1 \leq i \leq (c-1)$ and $j = 0$:

$$\begin{aligned}P_{i,0}(t+\delta t) =& P_{i,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-i\mu\delta t] \\ &+ P_{i-1,0}(t)[\lambda_1\delta t][1-\lambda_2\delta t][1-(i-1)\mu\delta t] \\ &+ P_{i+1,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](i+1)\mu\delta t \\ &+ P_{i,1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](i+1)\mu\delta t + o(\delta t)\end{aligned}$$

(3) For $i = 0$ and $1 \leq j \leq (c-1)$:

$$\begin{aligned}P_{0,j}(t+\delta t) =& P_{0,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-j\mu\delta t] \\ &+ P_{0,j-1}(t)[1-\lambda_1\delta t][\lambda_2\delta t][1-(j-1)\mu]\delta t] \\ &+ P_{0,j+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](j+1)\mu\delta t \\ &+ P_{1,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](j+1)\mu\delta t + o(\delta t)\end{aligned}$$

(4) For $i + j \leq (c-1)$ and $i, j \neq 0$:

$$\begin{aligned}P_{i,j}(t+\delta t) =& P_{i,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-2(i+j)\mu\delta t] \\ &+ P_{i-1,j}(t)[\lambda_1\delta t][1-\lambda_2\delta t][1-(i+j-1)\mu\delta t] \\ &+ P_{i,j-1}(t)[1-\lambda_1\delta t][\lambda_2\delta t][1-(i+j-1)\mu\delta t] \\ &+ P_{i+1,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](i+j+1)\mu\delta t \\ &+ P_{i,j+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](i+j+1)\mu\delta t + o(\delta t)\end{aligned}$$

(5) For $i = 0$ and $j = c$:

$$\begin{aligned}P_{0,c}(t+\delta t) =& P_{0,c}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-c\mu\delta t]\\ &+ P_{0,c-1}(t)[1-\lambda_1\delta t][\lambda_2\delta t][1-(c-1)\mu\delta t]\\ &+ P_{1,c}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\delta t\\ &+ P_{0,c+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\delta t + o(\delta t)\end{aligned}$$

(6) For $i = c$ and $j = 0$:

$$\begin{aligned}P_{c,0}(t+\delta t) =& P_{0,c}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-c\mu\delta t]\\ &+ P_{c-1,0}(t)[\lambda_1\delta t][1-\lambda_2\delta t][1-(c-1)\mu\delta t]\\ &+ P_{c+1,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\delta t\\ &+ P_{c,1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\delta t + o(\delta t)\end{aligned}$$

(7) For $i + j = c$ and $i, j \neq 0$:

$$\begin{aligned}P_{i,j}(t+\delta t) =& P_{i,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-2c\mu\delta t]\\ &+ P_{i-1,j}(t)[\lambda_1\delta t][1-\lambda_2\delta t][1-(i+j-1)\mu\delta t]\\ &+ P_{i,j-1}(t)[1-\lambda_1\delta t][\lambda_2\delta t][1-(i+j-1)\mu\delta t]\\ &+ P_{i+1,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\\ &+ P_{i,j+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu + o(\delta t)\end{aligned}$$

(8) For $c + 1 \leq i \leq c + m - 1$ and $j = 0$:

$$\begin{aligned}P_{i,0}(t+\delta t) =& P_{i,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-c\mu\delta t]\\ &+ P_{i-1,0}(t)[\lambda_1\delta t][1-\lambda_2\delta t][1-c\mu\delta t]\\ &+ P_{i+1,0}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\delta t\\ &+ P_{i,1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\delta t + o(\delta t)\end{aligned}$$

(9) For $c + 1 \leq j \leq c + m - 1$ and $i = 0$:

$$\begin{aligned}P_{0,j}(t+\delta t) =& P_{0,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-c\mu\delta t]\\ &+ P_{0,j-1}(t)[1-\lambda_1\delta t][\lambda_2\delta t][1-c\mu\delta t]\\ &+ P_{1,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\delta t\\ &+ P_{0,j+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\delta t + o(\delta t)\end{aligned}$$

(10) For $c+1 \leq i+j \leq c+m-1$ and $i,j \neq 0$:

$$\begin{aligned} P_{i,j}(t+\delta t) =& P_{i,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-2c\mu\delta t] \\ &+ P_{i-1,j}(t)\lambda_1\delta t[1-\lambda_2\delta t][1-c\mu\delta t] \\ &+ P_{i,j-1}(t)[1-\lambda_1\delta t][\lambda_2\delta t][1-c\mu\delta t] \\ &+ P_{i+1,j}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\delta t \\ &+ P_{i,j+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]c\mu\delta t + o(\delta t) \end{aligned}$$

(11) For $i=0$ and $j=c+m$:

$$\begin{aligned} P_{0,c+m}(t+\delta t) =& P_{0,c+m}(t)[1-c\mu\delta t] \\ &+ P_{0,c+m-1}(t)[1-\lambda_1\delta t][\lambda_2\delta t][1-c\mu\delta t] + o(\delta t) \end{aligned}$$

(12) For $i=c+m$ and $j=0$:

$$\begin{aligned} P_{c+m,0}(t+\delta t) =& P_{c+m,0}(t)[1-c\mu\delta t] \\ &+ P_{c+m-1,0}(t)[\lambda_1\delta t][1-\lambda_2\delta t][1-c\mu\delta t] + o(\delta t) \end{aligned}$$

(13) For $i+j=c+m$ and $i,j \neq 0$:

$$\begin{aligned} P_{i,j}(t+\delta t) =& P_{i,j}(t)[1-2c\mu\delta t] \\ &+ P_{i-1,j}(t)[\lambda_1\delta t][1-\lambda_2\delta t][1-c\mu\delta t] \\ &+ P_{i,j-1}(t)[1-\lambda_1\delta t]\lambda_2\delta t[1-c\mu\delta t] + o(\delta t) \end{aligned}$$

The steady-state equations for each of the 13 categories may be written in the following form:

(1) For $i=j=0$:

$$(\lambda_1+\lambda_2)P_{0,0} = \mu P_{1,0} + \mu P_{0,1}$$

(2) For $1 \leq i \leq (c-1)$ and $j=0$:

$$(\lambda_1+\lambda_2+i\mu)P_{i,0} = \lambda_1 P_{i-1,0} + (i+1)\mu P_{i+1,0} + (i+1)\mu P_{i,1}$$

(3) For $i=0$ and $1 \leq j \leq (c-1)$:

$$(\lambda_1+\lambda_2+j\mu)P_{0,j} = \lambda_2 P_{0,j-1} + (j+1)\mu P_{0,j+1} + (j+1)\mu P_{1,j}$$

(4) For $i + j \leq (c - 1)$ and $i, j \neq 0$:

$$[\lambda_1 + \lambda_2 + 2(i + j)\mu]P_{i,j} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + (i + j + 1)\mu P_{i+1,j} + (i + j + 1)\mu P_{i,j+1}$$

(5) For $i = 0$ and $j = c$:

$$(\lambda_1 + \lambda_2 + c\mu)P_{0,c} = \lambda_2 P_{0,c-1} + c\mu P_{1,c} + c\mu P_{0,c+1}$$

(6) For $i = c$ and $j = 0$:

$$(\lambda_1 + \lambda_2 + c\mu)P_{c,0} = \lambda_1 P_{c-1,0} + c\mu P_{c+1,0} + c\mu P_{c,1}$$

(7) For $i + j = c$ and $i, j \neq 0$:

$$(\lambda_1 + \lambda_2 + 2c\mu)P_{i,j} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + c\mu P_{i+1,j} + c\mu P_{i,j+1}$$

(8) For $c + 1 \leq i \leq c + m - 1$ and $j = 0$:

$$(\lambda_1 + \lambda_2 + c\mu)P_{i,0} = \lambda_1 P_{i-1,0} + c\mu P_{i+1,0} + c\mu P_{i,1}$$

(9) For $c + 1 \leq j \leq c + m - 1$ and $i = 0$:

$$(\lambda_1 + \lambda_2 + c\mu)P_{0,j} = \lambda_2 P_{0,j-1} + c\mu P_{1,j} + c\mu P_{0,j+1}$$

(10) For $c + 1 \leq i + j \leq c + m - 1$ and $i, j \neq 0$:

$$(\lambda_1 + \lambda_2 + 2c\mu)P_{i,j} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + c\mu P_{i+1,j} + c\mu P_{i,j+1}$$

(11) For $i = 0$ and $j = c + m$:

$$c\mu P_{0,c+m} = \lambda_2 P_{0,c+m-1}$$

(12) For $i = c + m$ and $j = 0$:

$$c\mu P_{c+m,0} = \lambda_1 P_{c+m-1,0}$$

(13) For $i + j = c + m$ and $i, j \neq 0$:

$$2c\mu P_{i,j} = \lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} \tag{3.6}$$

Theorem 3.3.2. *The steady-state probability that there are i emergency and j elective patients present with different arrival rates (λ_1 and λ_2) and combined service rate μ in the system where queue of size m is allowed is given by:*

$$P_{i,j} = \begin{cases} \frac{1}{(i+j)!}{\theta_1}^i{\theta_2}^j P_{0,0} & \text{if } 1 \leq i+j \leq c \\ \frac{1}{c^{i+j-c}c!}{\theta_1}^i{\theta_2}^j P_{0,0} & \text{if } c \leq i+j \leq c+m \end{cases}$$

where

$$\mu = \frac{\lambda}{\theta_1 + \theta_2}, \qquad \lambda = \lambda_1 + \lambda_2$$

$$\theta_k = \frac{\lambda_k}{\mu_k}, \qquad k = 1, 2$$

Proof.
The proof of this result is relatively straightforward, using an inductive approach. The required algebraic manipulation for each of the steady-state equations in Formula 3.6 is given in Appendix A. □

$P_{0,0}$ is calculated in the usual way using the fact that $\sum_{i=0}^{c+m}\sum_{j=0}^{c-i} P_{i-j,j} = 1$.

$$P_{0,0} = \frac{1}{\sum_{i=0}^{c}\sum_{j=0}^{c-i}\frac{1}{i!}{\theta_1}^{i-j}{\theta_2}^j + \sum_{i=c+1}^{c+m}\sum_{j=0}^{c+m-i}\frac{1}{c^{i-c}c!}{\theta_1}^{i-j}{\theta_2}^j}$$

3.3.4 Connection Between $M_2/M_2/c/c/FIFO$ and $M/M/c/c/FIFO$

Recall Section 3.3.2 where two independent admission streams were combined into a single input and similarly two service rates combined into a single service rate. There is a clear connection between $M_2/M_2/c/c/FIFO$ and $M/M/c/c/FIFO$ type of queueing model. A simple derivation of the steady state probabilities for the $M/M/c/c/FIFO$ queue with two different arrival streams and two different service rates will be presented in this section.

Consider the $M/M/c/c$ queue with the following parameters:

λ	the arrival rate
μ	the service rate
c	the number of servers
$\theta = \frac{\lambda}{\mu}$	the service intensity
P_n	the probability of having n beds occupied

The following steady state probabilities are known (Stewart, 2009 [152]) to be:

$$P_n = \frac{1}{n!}\theta^n P_0 \tag{3.7}$$

Consider the slight generalisation of this model, so that two service types are allowed. For each service type an arrival rate is as follows: λ_1 for type I customers, λ_2 for type II customers and a service rates: μ_1 for type I customers and μ_2 for type II customers. Let $P_{i,j}$ denote the probability of having i type I customers and j type II customers in the system.

Recall, as shown in Section 3.3.2, the mean arrival rate is given by $\lambda = \lambda_1 + \lambda_2$, and the mean service rate:

$$\mu = \frac{\lambda}{\theta_1 + \theta_2}$$

The steady state probabilities can be obtained for the total number of customers in the system using Formula 3.7 with $\theta = \theta_1 + \theta_2$ (where $\theta_1 = \frac{\lambda_1}{\mu_1}$ and $\theta_2 = \frac{\lambda_2}{\mu_2}$).

Using the definition of conditional probability:

$$P_{i,j|i+j} = \frac{P_{(i,j)\cap(i+j)}}{P_{i+j}} = \frac{P_{i,j}}{P_{i+j}} \tag{3.8}$$

Rearranging Formula 3.8:

$$P_{i,j} = P_{i+j} \times P_{i,j|i+j} \tag{3.9}$$

where $P_{i,j|i+j}$ is the probability of having i type I customers and j type II customers given $i + j$ customers in the system.

As discussed previously, P_{i+j} is given by Formula 3.7. Thus an expression for $P_{i,j|i+j}$ is required. Let p denote the probability that a server is used by a type I customer and q denote the probability that a server is used by a type II customer. Then:

$$P_{i,j|i+j} = \binom{i+j}{i} p^i q^j = \binom{i+j}{j} p^i q^j \tag{3.10}$$

as ordering of servers is not important.

Both p and q are functions of $\lambda_1, \lambda_2, \mu_1, \mu_2$ so that $p = p(\lambda_1, \lambda_2, \mu_1, \mu_2)$ and $q = q(\lambda_1, \lambda_2, \mu_1, \mu_2)$. By rescaling, the problem can be reduced to the case $\mu_1 = \mu_2 = 1$, i.e. occupancy of the bed by emergency and elective patient is the same in both cases, therefore:

$$p(\lambda_1, \lambda_2, \mu_1, \mu_2) = p(\widehat{\lambda_1}, \widehat{\lambda_2}, 1, 1)$$

Thus, the expressions for $\widehat{\lambda_1}, \widehat{\lambda_2}$ are required. Type I customers arrive at rate λ_1 and stay in the system for $\frac{1}{\mu_1}$ length of time. If their length of stay was now 1, the arrival rate would have to be increased by a factor of $\frac{1}{\mu_1}$. Giving $\widehat{\lambda_1} = \frac{\lambda_1}{\mu_1}$, or equivalently $\widehat{\lambda_2} = \frac{\lambda_2}{\mu_2}$.
Therefore:

$$p = \frac{\frac{\lambda_1}{\mu_1}}{\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}} \qquad \text{and} \qquad q = \frac{\frac{\lambda_2}{\mu_2}}{\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}} \tag{3.11}$$

Combining Equations 3.10 and 3.11, gives:

$$\begin{aligned} P_{i,j|i+j} &= \binom{i+j}{i} \left(\frac{\frac{\lambda_1}{\mu_1}}{\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}} \right)^i \left(\frac{\frac{\lambda_2}{\mu_2}}{\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}} \right)^j \\ &= \binom{i+j}{i} \frac{(\theta_1)^i}{(\theta_1+\theta_2)^i} \frac{(\theta_2)^j}{(\theta_1+\theta_2)^j} \\ &= \binom{i+j}{i} \frac{(\theta_1)^i (\theta_2)^j}{(\theta_1+\theta_2)^{i+j}} \end{aligned} \tag{3.12}$$

Therefore Formula 3.9 gives:

$$\begin{aligned} P_{i,j} &= P_{i+j} \times P_{i,j|i+j} \\ &= \frac{1}{(i+j)!}(\theta_1+\theta_2)^{i+j} P_0 \frac{(i+j)!}{i!j!} \frac{(\theta_1)^i(\theta_2)^j}{(\theta_1+\theta_2)^{i+j}} \\ &= \frac{1}{i!j!}(\theta_1)^i(\theta_2)^j P_0 \qquad \text{(where } P_0 = P_{0,0}\text{)} \end{aligned} \tag{3.13}$$

as required.

3.3.5 The Multi-Class $M_k/M_k/c/c/FIFO$ Queue

This section is a generalisation of Section 3.3.4, where only two service types were allowed. This section will consider $k \in \mathbb{Z}$ service types. Each service type has an arrival rate λ_i and a service rate μ_i for $i \in [k]$. $P_{j_1,\ldots,j_k}$ denotes the probability of having $j_i \in \mathbb{Z}$ customers of type i in the system for $i \in [k]$.

Given the above parameters the mean arrival rate is: $\lambda = \sum_{i=1}^{k} \lambda_i$. The mean service rate μ can be obtained by observing that a mean service time $\frac{1}{\mu_i}$ is by a customer with probability $\frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i}$ and so the mean service time is:

$$\begin{aligned}
\frac{1}{\mu} &= \sum_{k=1}^{n} \frac{\lambda_k}{\mu_k \sum_{i=1}^{k} \lambda_i} \\
&= \frac{\sum_{k=1}^{n} \frac{\lambda_k}{\mu_k \sum_{i=1}^{k} \lambda_i}}{\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{k} \lambda_i}} && \text{(as a weighted mean)} \\
&= \frac{\sum_{i=1}^{k} \frac{\lambda_i}{\mu_i}}{\sum_{i=1}^{k} \lambda_i} \\
&= \frac{\sum_{i=1}^{k} \theta_i}{\lambda} && \text{(letting } \theta_i = \frac{\lambda_i}{\mu_i} \text{ for } i \in [k]\text{)} \qquad (3.14)
\end{aligned}$$

this gives:

$$\mu = \frac{\lambda}{\sum_{i=1}^{k} \theta_i}$$

and the steady state probabilities for the total number of customers in the system can be obtained using Formula 3.7 with $\theta = \sum_{i=1}^{k} \theta_i$.

To obtain the $P_{j_1,\ldots,j_k}$ Formula 3.8 gives:

$$P_{j_1,\ldots,j_k} = P_{j_1+\cdots+j_k} \times P_{j_1,\ldots,j_k|j_1+\cdots+j_k} \tag{3.15}$$

As discussed before, $P_{j_1+\cdots+j_k}$ is given by Formula 3.7. Thus an expression for $P_{j_1,\ldots,j_k|j_1+\cdots+j_k}$ needs to be obtained. Let p_i denote the probability that a server is occupied by a customer of type i then:

$$P_{j_1,\ldots,j_k|j_1+\cdots+j_k} = \binom{j_1 + \cdots + j_k}{j_1, \ldots, j_k} \prod_{i=1}^{k} p_i^{j_i} \tag{3.16}$$

as ordering of servers is not important and $\binom{j_1+\cdots+j_k}{j_1,\ldots,j_k}$ is the multinomial coefficient of order k.

Since $p_i = p_i(\lambda_1, \ldots, \lambda_k, \mu_1, \ldots, \mu_k)$ it is straightforward to obtain:

$$p_i(\lambda_1, \ldots, \lambda_k, 1, \ldots, 1) = \frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i}$$

Without loss of generality λ_i can be rescaled to get:

$$p_i(\lambda_i, \ldots, \lambda_k, \mu_1, \ldots, \mu_k) = p_i(\theta_1, \ldots, \theta_k, 1, \ldots, 1)$$

and so:

$$p_i(\lambda_i, \ldots, \lambda_k, \mu_1, \ldots, \mu_k) = \frac{\theta_i}{\sum_{i=1}^{k} \theta_i}$$

Combining Equations (3.7, 3.14 - 3.16) gives:

$$\begin{aligned} P_{j_1,\ldots,j_k} &= \frac{1}{\left(\sum_{i=1}^{k} j_i\right)!} \left(\sum_{i=1}^{k} \theta_i\right)^{\sum_{i=1}^{k} j_i} P_0 \binom{\sum_{i=1}^{k} j_i}{j_1, \ldots, j_k} \prod_{i=1}^{k} \left(\frac{\theta_i}{\sum_{i=1}^{k} \theta_i}\right)^{j_i} \\ &= \frac{1}{\left(\sum_{i=1}^{k} j_i\right)!} \left(\sum_{i=1}^{k} \theta_i\right)^{\sum_{i=1}^{k} j_i} \frac{\left(\sum_{i=1}^{k} j_i\right)!}{\prod_{i=1}^{k} j_i!} \frac{\prod_{i=1}^{k} {\theta_i}^{j_i}}{\left(\sum_{i=1}^{k} \theta_i\right)^{\sum_{i=1}^{k} j_i}} P_0 \\ &= \frac{1}{\prod_{i=1}^{k} j_i!} \prod_{i=1}^{k} {\theta_i}^{j_i} P_0 \end{aligned} \tag{3.17}$$

(where $P_0 = P_{0,\ldots,0}$) as required.

Therefore, the probability of having $j_i \in \mathbb{Z}$ customers of type i in the system for $i \in [k]$ having an arrival rate λ_i and a service rate μ_i for $i \in [k]$ is given by:

$$P_{j_1,\ldots,j_k} = \frac{1}{\prod_{i=1}^{k} j_i!} \prod_{i=1}^{k} {\theta_i}^{j_i} P_0$$

3.3.6 Queueing Model with Cut-off

Using the Formula 3.5 for the combined service rate μ, consideration is given to reducing variation by not allowing any elective admissions when the bed occupancy reaches a pre-determined cut-off point. The cut-off point is denoted by k. Consider the case when there is a restriction on the number of patients allowed in the system. When considering the CCU, a restriction may well be placed on the queue, with patients being referred elsewhere when bed occupancy levels become too high. The differential-difference equations were set up to describe the system with c service channels available. The difference between the differential-difference equations derived earlier (Formula 3.1) is that the formula for P_n, the probability of the number of beds occupied is considered, not $P_{i,j}$ as before. These equations can be categorised into five types as shown below.

(1) For $n = 0$:

$$\begin{aligned} P_0(t + \delta t) =& P_0(t)[1 - \lambda_1 \delta t][1 - \lambda_2 \delta t] \\ &+ P_1(t)[1 - \lambda_1 \delta t][1 - \lambda_2 \delta t]\mu \delta t + o(\delta t) \end{aligned}$$

(2) For $1 \leq n \leq (k-1)$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)[1-(\lambda_1\delta t+\lambda_2\delta t)][1-n\mu\delta t] \\ &+ P_{n-1}(t)[\lambda_1\delta t+\lambda_2\delta t][1-(n-1)\mu\delta t] \\ &+ P_{n+1}(t)[1-\lambda_1\delta t](n+1)\mu\delta t + o(\delta t) \end{aligned}$$

(3) For $n = k$:

$$\begin{aligned} P_k(t+\delta t) =& P_k(t)[1-\lambda_1\delta t][1-k\mu\delta t] \\ &+ P_{k-1}(t)[\lambda_1\delta t+\lambda_2\delta t][1-(k-1)\mu\delta t] \\ &+ P_{k+1}(t)[1-\lambda_1\delta t](k+1)\mu\delta t + o(\delta t) \end{aligned}$$

(4) For $(k+1) \leq n \leq c-1$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)[1-\lambda_1\delta t][1-n\mu\delta t] \\ &+ P_{n-1}(t)[\lambda_1\delta t][1-(n-1)\mu\delta t] \\ &+ P_{n+1}(t)[(n+1)\mu\delta t] + o(\delta t) \end{aligned}$$

(5) For $n = c$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)[1-n\mu\delta t] \\ &+ P_{n-1}(t)[\lambda_1\delta t][1-(n-1)\mu\delta t] + o(\delta t) \end{aligned} \tag{3.18}$$

The steady-state equations are given below.

(1) For $n = 0$:

$$\mu P_1 = (\lambda_1+\lambda_2)P_0$$

(2) For $1 \leq n \leq (k-1)$:

$$(\lambda_1+\lambda_2+n\mu)P_n = (\lambda_1+\lambda_2)P_{n-1} + (n+1)\mu P_{n+1}$$

(3) For $n = k$:

$$(\lambda_1+n\mu)P_n = (\lambda_1+\lambda_2)P_{n-1} + (n+1)\mu P_{n+1}$$

(4) For $(k+1) \leq n \leq c-1$:

$$(\lambda_1 + n\mu)P_n = \lambda_1 P_{n-1} + (n+1)\mu P_{n+1}$$

(5) For $n = c$:

$$n\mu P_n = \lambda_1 P_{n-1} \tag{3.19}$$

Theorem 3.3.3. *The steady-state bed occupancy probabilities, with a restriction on the number of customers allowed in the system, with a cut-off point k is given by:*

$$P_n = \begin{cases} \frac{1}{n!}\theta^n P_0, & \text{if } n \leq k \\ \frac{1}{n!}{\theta_1}^{n-k}\theta^k P_0, & \text{if } n > k \end{cases}$$

where

$$\theta = \theta_1 + \theta_2 = \frac{\lambda_1 + \lambda_2}{\mu}$$

and

$$\theta_1 = \frac{\lambda_1}{\mu}; \qquad \theta_2 = \frac{\lambda_2}{\mu}$$

Proof.
The proof is similar to that of Theorem 3.3.1 and is again done using an inductive approach. The required algebraic manipulation for each of the Equations 3.19 is given in Appendix B. □

P_0 was calculated in the usual way using the fact that $\sum_{n=0}^{c} P_n = 1$.

$$P_0 = \frac{1}{1 + \sum_{r=1}^{k} \frac{1}{r!}\theta^r + \theta^k \sum_{r=k+1}^{c} \frac{1}{r!}(\theta_1)^{r-k}}$$

A program written in Visual Basic calculates the probabilities of each bed occupancy, P_n, where $n = 0, 1, \ldots, 29$, for different values of the cut-off parameter k. The arrival and service rate parameters used are given in Table 3.4. Note that elective service rate μ_2 is not included since before the cut-off both patients' types are served, so the overall value of μ is used, and after cut-off only emergency patients are served with rate μ_1.

The results of the model are explored for different values of cut-off points, from 19 to 24. Figure 3.13 illustrates the comparison of observed bed occupancy probabilities with analytical results for the cut-off chosen to be at 24 beds. As expected, high bed occupancy probabilities are no longer overestimated.

Table 3.4: Parameter values

Parameter	Value
λ_1	2.6081
λ_2	1.2231
μ_1	0.1406
μ	0.1718

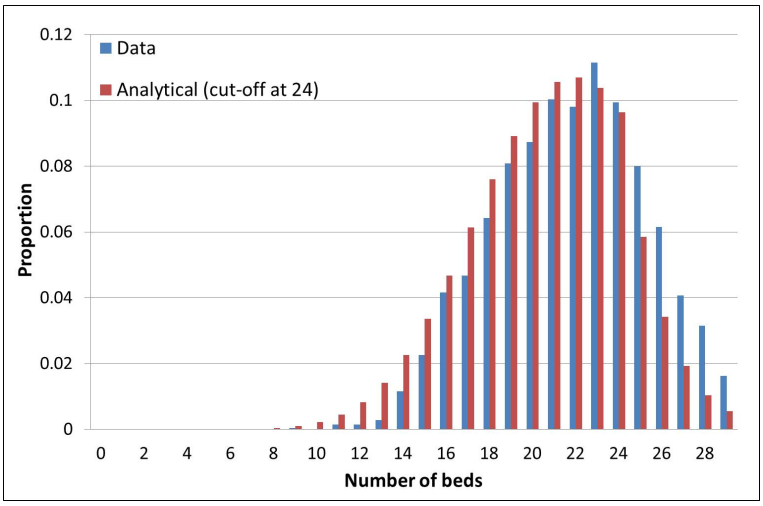


Figure 3.13: Comparison of analytical results with the cut-off point at 24 and the original data

The effect of that 'what if' scenario is highly influential. The measures that are again examined are: the mean and the standard deviation of the bed occupancy. These measures found for cut-off points between 19 and 24 are presented in Table 3.5.

Table 3.5: Mean and standard deviation of the bed occupancy from the model for different values of cut-off point

Cut-off point	Mean	Standard Deviation
19	18.57	3.12
20	19.03	3.17
21	19.47	3.24
22	19.89	3.34
23	20.27	3.45
24	20.62	3.57

Unsurprisingly, as the cut-off point increases, the mean and standard deviation of the bed occupancy rise.

3.3.7 Queueing Model with Cut-off and Extra Admissions

It is now necessary to consider the implications of the different changes in the mode of operation of the CCU. Recall that the main method of control over the rate of admissions to the CCU is by changing the admission rates of elective patients. The relationship between elective admissions and bed occupancy levels has been studied before with the purpose of developing improved elective admissions schedules (Gallivan and Utley, 2005 [56]).

The number of patients admitted (throughput) averaged 1406 per year and there was a considerable degree of variation in the bed occupancy levels with a standard deviation of 3.57 beds. This section will investigate a further model that allows for the increase of the average number of elective admissions whenever there are between a and b beds occupied. That is, whenever there appears to be sufficient spare bed capacity, then it is suggested that extra (here two) elective patients be admitted. No elective admissions are allowed when there are more than k (cut-off point) beds occupied. New differential-difference equations were set up again to describe the system. The system works in the following way:

- If there are more than a and less than b beds occupied, the number of elective admissions is increased by 2. The new mean elective arrival rate is denoted by λ_2'.
- If there are more than b beds occupied and less than k, the number of elective admissions is not changed.
- If there are more than k beds occupied, the number of elective admissions is set to zero.

The system with c service channels available is described by the set of the differential-difference equations given in the following Equations 3.20.

(1) For $n = 0$:

$$\begin{aligned} P_0(t+\delta t) =& P_0(t)[1-\lambda_1\delta t][1-\lambda_2\delta t] \\ &+ P_1(t)[1-\lambda_1\delta t][1-\lambda_2\delta t]\mu\delta t + o(\delta t) \end{aligned}$$

(2) For $1 \leq n \leq (a-1)$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-n\mu\delta t] \\ &+ P_{n-1}(t)[\lambda_1\delta t+\lambda_2\delta t][1-(n-1)\mu\delta t] \\ &+ P_{n+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](n+1)\mu\delta t + o(\delta t) \end{aligned}$$

(3) For $n = a$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)[1-\lambda_1\delta t][1-\lambda_2'\delta t][1-n\mu\delta t] \\ &+ P_{n-1}(t)[\lambda_1\delta t+\lambda_2\delta t][1-(n-1)\mu\delta t] \\ &+ P_{n+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](n+1)\mu\delta t + o(\delta t) \end{aligned}$$

(4) For $(a+1) \leq n \leq b$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)[1-\lambda_1\delta t][1-\lambda_2'\delta t][1-n\mu\delta t] \\ &+ P_{n-1}(t)[\lambda_1\delta t+\lambda_2'\delta t][1-(n-1)\mu\delta t] \\ &+ P_{n+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](n+1)\mu\delta t + o(\delta t) \end{aligned}$$

(5) For $n = b+1$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-n\mu\delta t] \\ &+ P_{n-1}(t)[\lambda_1\delta t+\lambda_2'\delta t][1-(n-1)\mu\delta t] \\ &+ P_{n+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](n+1)\mu\delta t + o(\delta t) \end{aligned}$$

(6) For $(b+2) \leq n \leq (k-1)$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)[1-\lambda_1\delta t][1-\lambda_2\delta t][1-n\mu\delta t] \\ &+ P_{n-1}(t)[\lambda_1\delta t+\lambda_2\delta t][1-(n-1)\mu\delta t] \\ &+ P_{n+1}(t)[1-\lambda_1\delta t][1-\lambda_2\delta t](n+1)\mu\delta t + o(\delta t) \end{aligned}$$

(7) For $n = k$:

$$\begin{aligned} P_k(t+\delta t) =& P_k(t)[1-\lambda_1\delta t][1-k\mu\delta t] \\ &+ P_{k-1}(t)[\lambda_1\delta t+\lambda_2\delta t][1-(k-1)\mu\delta t] \\ &+ P_{k+1}(t)[1-\lambda_1\delta t](k+1)\mu\delta t + o(\delta t) \end{aligned}$$

(8) For $(k+1) \leq n \leq (c-1)$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)[1-\lambda_1\delta t][1-n\mu\delta t] \\ &+ P_{n-1}(t)[\lambda_1\delta t][1-(n-1)\mu\delta t] \\ &+ P_{n+1}(t)[1-\lambda_1\delta t][(n+1)\mu\delta t] + o(\delta t) \end{aligned}$$

(9) For $n = c$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)[1 - n\mu\delta t] \\ &+ P_{n-1}(t)[\lambda_1\delta t][1-(n-1)\mu\delta t] + o(\delta t) \end{aligned} \tag{3.20}$$

The steady-state equations may be written in the form shown below in Equation 3.21.

(1) For $n = 0$:

$$\mu P_1 = (\lambda_1 + \lambda_2)P_0$$

(2) For $1 \leq n \leq (a-1)$:

$$(\lambda_1 + \lambda_2 + n\mu)P_n = (\lambda_1 + \lambda_2)P_{n-1} + (n+1)\mu P_{n+1}$$

(3) For $n = a$:

$$(\lambda_1 + \lambda_2{}' + n\mu)P_n = (\lambda_1 + \lambda_2)P_{n-1} + (n+1)\mu P_{n+1}$$

(4) For $(a+1) \leq n \leq b$:

$$(\lambda_1 + \lambda_2{}' + n\mu)P_n = (\lambda_1 + \lambda_2{}')P_{n-1} + (n+1)\mu P_{n+1}$$

(5) For $n = b+1$:

$$(\lambda_1 + \lambda_2 + n\mu)P_n = (\lambda_1 + \lambda_2')P_{n-1} + (n+1)\mu P_{n+1}$$

(6) For $(b+2) \leq n \leq (k-1)$:

$$(\lambda_1 + \lambda_2 + n\mu)P_n = (\lambda_1 + \lambda_2)P_{n-1} + (n+1)\mu P_{n+1}$$

(7) For $n = k$:

$$(\lambda_1 + n\mu)P_n = (\lambda_1 + \lambda_2)P_{n-1} + (n+1)\mu P_{n+1}$$

(8) For $(k+1) \leq n \leq (c-1)$:

$$(\lambda_1 + n\mu)P_n = \lambda_1 P_{n-1} + (n+1)\mu P_{n+1}$$

(9) For $n = c$:

$$n\mu P_n = \lambda_1 P_{n-1} \tag{3.21}$$

Theorem 3.3.4. *The steady-state bed occupancy probabilities with cut-off point k and increase of elective admissions for bed occupancy levels between a and b is given by:*

$$P_n = \begin{cases} \frac{1}{n!}\theta^n P_0, & \text{if } 0 \leq n \leq a \\ \frac{1}{n!}\theta^a(\theta_1 + \theta_2')^{n-a} P_0, & \text{if } (a+1) \leq n \leq (b+1) \\ \frac{1}{n!}\theta^{n-b+a-1}(\theta_1 + \theta_2')^{b-a+1} P_0, & \text{if } (b+2) \leq n \leq k \\ \frac{1}{n!}\theta^{k-1-b+a}{\theta_1}^{n-k}(\theta_1 + \theta_2')^{b-a+1} P_0, & \text{if } (k+1) \leq n \leq c \end{cases}$$

where

$$\theta_1 = \frac{\lambda_1}{\mu} \qquad \textit{and} \qquad \theta_2 = \frac{\lambda_2}{\mu} \qquad \textit{and} \qquad \theta_2' = \frac{\lambda_2'}{\mu}$$

and

$$\theta = \theta_1 + \theta_2 = \frac{\lambda_1 + \lambda_2}{\mu}$$

$$P_0 = \frac{1}{\sum_{r=0}^{a} \frac{1}{r!}\theta^r + \theta^a \sum_{r=a+1}^{b+1} \frac{1}{r!}(\theta_1 + \theta_2')^{r-a} + (\theta_1 + \theta_2')^{b-a+1} \sum_{r=b+2}^{k} \frac{1}{r!}\theta^{r-b+a-1} + \theta^{k-1-b+a}(\theta_1 + \theta_2')^{r-b+a-1} \sum_{r=k+1}^{c} \frac{1}{r!}\theta_1^{r-k}}$$

Proof.
The proof is similar to that of Theorem 3.3.1 and again utilises an inductive approach. The required algebraic manipulation for each of the Equations 3.21 is given in Appendix C. □

The effect of admitting more elective patients at non-busy times is analysed and the analytical results obtained are compared with the original data. A program written in Visual Basic calculates the probabilities for each bed occupancy P_n, where $n = 0, 1, ..., 29$. Values for the parameters a and b can easily be changed. For purpose of this thesis results are obtained for the parameters given in Table 3.6.

Table 3.6: Parameter values

Parameter	**Value**
a	1
b	22
k (cut-off)	24

The result of the analytical model with cut-off at 24 and two extra elective patients when bed occupancy is between 1 and 22 is presented in Figure 3.14. It shows higher probabilities of mid-value

bed occupancy levels and small probabilities of low and high bed occupancy levels. As an objective of this project was to decrease variation of the system, to confirm the effect of that scenario the mean and the standard deviation are investigated. The mean bed occupancy increases to 22.15, a 2% increase and the standard deviation was reduced to 2.81, a significant decrease of 20.2%.

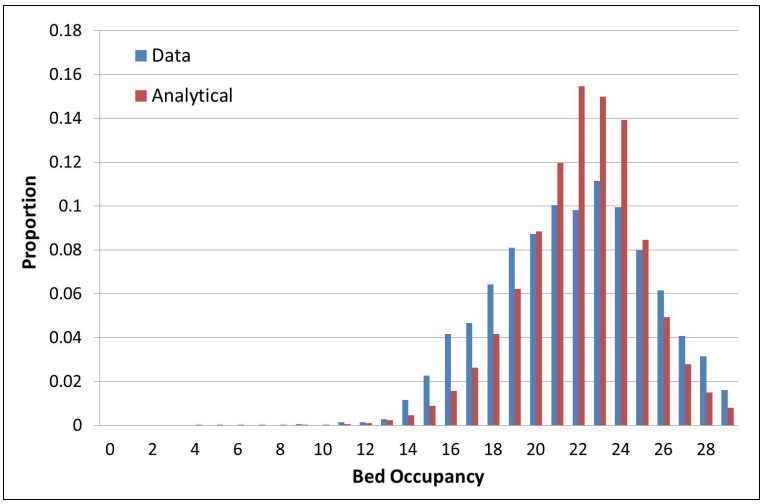


Figure 3.14: Reduction in variability and increase in throughput

As mentioned before, throughput of patients is of a great interest; it is increased to 1701 patients per year, again significant increase of 21%. Thus, a relatively minor increase in admissions of elective patients at non-busy times shows a marked improvement in variability and throughput.

A sensitivity analysis is undertaken to check how the throughput changes for different values of $a \in \{1, 2, \ldots, 22\}$ and $b \in \{a+1, \ldots, 23\}$. It can be concluded that the bigger the difference between these values the higher the throughput, but also the throughput is constant if $a \leq 8$ for any given value of b.

If the elective patient increase is changed from two to four per day at non-busy times, then the corresponding reduction in standard deviation is 31.6% (2.41), and the increase in throughput is 39% (1958).

3.4 Conclusions

This chapter started with a simple simulation of the CCU, followed by two 'what if' scenarios. The first scenario investigated the effect of not allowing any elective admissions when bed occupancy reached a pre-determined cut-off level. The second scenario was an extension to the first one, where extra elective patients were admitted at non-busy times.

Section 3.3 concentrated on setting up and solving the relevant queueing equations. In Section 3.3.2 a loss queueing model was developed to find probabilities of having multiple customer types with different arrival and service rates. In Section 3.3.3 a queueing model with multiple arrival rates and a combined service rate was developed, where a small number of patients are allowed to queue.

Figure 3.15 summarizes the work undertaken in Section 3.3.2 and Section 3.3.4, which is a connection between two different probabilities. It has been shown that knowing one allows one to obtain the other.

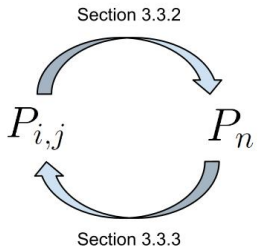


Figure 3.15: Diagram of connection between P_n and $P_{i,j}$

Section 3.3.6 considered a queueing model where restrictions are placed on admissions, with elective patients being referred elsewhere when bed occupancy levels reach a cut-off point. Finally, Section 3.3.7 extended the queueing model with a cut-off point, to model additional elective patients being admitted at non-busy periods.

The queueing models described in this section have not been developed in the literature and thus are considered to be original research contributions in this thesis.

Chapter 4

Further Applications of Mathematical Modelling at the Critical Care Unit in University Hospital of Wales

4.1 Introduction

This chapter will investigate further applications of mathematical modelling. Since the nature of elective admissions is time-dependant, Section 4.2 will give a non-stationary study of the bed occupancy. Section 4.3.1 will question of whether it is possible to predict the number of beds occupied at future days knowing today's day of the week and current bed occupancy. Finally, but very importantly from an operating point of view, Section 4.4 will provide a model that optimises the number of nurses to be employed per shift.

4.2 Time-Dependent Aspects

4.2.1 Literature Review

Figure 2.2 illustrates a weekly average profile of elective patients admissions to the CCU. Unlike emergency arrivals, elective demand is dependent on day of the week. The majority of elective admissions occur on Thursday with very few on the weekend.

Whilst much literature is devoted to the analysis of a service system with constant arrival and service times (Green and Kolesar, 1991 [66]; Pollaczek, 1934 [135]), most actual systems are subject to time-varying demand, where arrival rates and the number of servers vary throughout the period of operation. Computer systems, road traffics, telecommunications networks, banks, airports, toll booths and hospitals systems are just a few examples of facilities with time-varying demand pro-

cesses. The steady-state theory developed by Erlang was inadequate for time-dependent systems, and began numerical investigations into the behaviour of the system during a finite interval. The study of time-dependent queues remains a vibrant area of research. The process involved are far more complex and as a consequence more sophisticated mathematical procedures are necessary (Channouf *at al.*, 2007 [24]; Holcomb and Sharpe, 2007 [87]; Feldman *et al.*, 2008 [51]; Bekker and de Bruin, 2010 [13]; Caiado, 2010 [23]; Izady and Worthington, 2012 [92]). Analytical models for such situations are often intractable, but in addition to numerical approaches, approximation methods have been developed, that provide reliable results in suitable scenarios (Green *et al.*, 1991 [66]).

Several approximation methods have been proposed in the literature which use series of tractable stationary models to estimate the time-dependent nature of a system. These methods, however only give reliable results under certain conditions. They do not consider non-stationary and transient effects, so will only be accurate if the rate of change of the arrival rate relative to the throughput of the system is sufficiently low to allow the system to quickly achieve the steady state associated with any arrival rate (Utley and Worthington, 2011 [158]).

Two methods, which make use of compartmentalized steady-state models to find the minimum number of servers required to meet a desired service target in each planning period are the stationary independent period-by-period approach (SIPP) and the pointwise stationary approximation (PSA). Whilst variants of these methods have been developed in the literature over the last four decades (Kolesar *et al.*, 1975 [105]; Green and Kolesar, 1991 [66], 1997 [68]; Green *et al.*, 1991 [67], 2001 [69], 2006 [71]; Ingolfson *et al.*, 2007 [90]), an alternative method known as the modified-offered-load (MOL) has also been investigated (Massey and Whitt, 1994 [119], 1997 [120]; Ingolfsson *et al.*, 2007 [90]).

Whilst approximation methods provide instant solutions, the numerical approaches are able to offer solutions with a higher degree of accuracy at the expense of computational time.

In the literature, there exist four numerical approximations to the solution of differential equations; thus the $M(t)/M(t)/c(t)$ time-dependent equations can be solved using one of the following methods:

- Euler's method;
- Runge-Kutta method;
- the randomization method;
- the discrete time modelling (DTM) method;

Further details regarding the theoretical underpinnings of the numerical methods used to solve the time-dependent equations is provided in Gross and Harris, 1998 [77].

The randomization method (described in Grassmann, 1977 [63]) is a method to compute transient solutions of finite state continuous-time Markov chains. The method involves the constructions of an analogous discrete time Markov chain, where transitions occur according to an exponential distribution with the same parameter in every state. The method provided similar results as the Runge-Kutta (Ingolfsson *et al.*, 2007 [90]), but was more computationally efficient.

The DTM method produces accurate results at a much faster time than several approximation methods (Wall and Worthington, 2007 [165]). The approach uses discrete-time models to approximate the behaviour of continuous time queues by dividing the time of operation of the system into a set of non-overlapping intervals. The technique creates a transition matrix to take account of the various states that occur at each time step, and evaluate the probabilities associated with each state.

Euler's and Runge-Kutta methods are general approaches for solving ordinary differential equations. The Runge-Kutta approach provides solutions that are referred to as 'exact' since the only approximations required are the approximation of the infinite set of equations with a finite set, and those inherent in any numerical solution of ordinary differential equations. Euler's method, however has the advantage that it may be implemented to provide solutions at a quicker rate and does not require an ordinary differential equation solver (Izady, 2010 [91]).

Euler's method considers the slope of the tangent line to approximate the solution at each interval. It approximates the solution by evaluating the equations at a starting value, and then at steps separated by small time intervals δt (between which the solution is not expected to have changed greatly). Smaller step sizes generate solutions with higher accuracies, but this comes at a greater computational cost (Izady, 2010 [91]). This method is investigated in the form of case study in Section 4.2.2.

4.2.2 Time-Dependent Bed Utilisation

It was noted in Section 2.3.1 that the admission of patients undergoing elective surgery depends on day of the week. The importance of incorporating the time-dependent nature of these arrivals has been highlighted (Costa *et al.*, 2003 [35]). It was noted that the difference in the arrival rates during weekdays is not as visible as between weekdays and weekend. Therefore this section will adapt the mathematical model to take this factor into account and allow for different elective arrival rates for weekdays and weekend.

The differential-difference equations given in the set of Equations 3.1 are solved using a numerical iterative method (Euler's method), described in Section 4.2.1. A step-length of $\delta t = 0.01$ (hours) was used, and the process is initially run to produce bed occupancy levels for 1 week (168 hours). The iterative process is implemented in Visual Basic, linked to an Excel spreadsheet, which enables key variables to be altered quickly and easily. The Visual Basic program calculates the required probabilities, and outputs these to a table in Excel. The mean and the standard deviation of the number of beds occupied are also calculated and included in the results.

Consider the results evaluated when using the parameters obtained from the data and which are given in Table 4.1.

Table 4.1: Parameter values

Parameter	**Parameter name**	**Parameter Value**
c	Number of service channels (beds)	29
λ_1	Emergency arrival rate (per day)	2.6081
λ_2	Elective arrival rate at weekdays (per day)	1.5808
λ_2^W	Elective arrival rate at weekends (per day))	0.3542
μ_1	Emergency service rate (days)	0.1411
μ_2	Elective service rate (days)	0.3262
δt	Time increment (hours)	0.01

The appropriate values of λ_2, are used for the first 120 hours of the week (i.e. Monday to Friday), and then the reduced value for λ_2^W are used over the period 120-168 hours (Saturday and Sunday).

Initially (at time $t = 0$), P_0, the probability of having an empty system is set equal to 1 with all other initial probabilities (P_n, where $n = 1, \ldots, 29$) being set to 0. In this way, it is assured that the initial assignment of probabilities sums to one. This process is repeated and a new set of probabilities are generated with every increment of δt, using the probabilities at time t to calculate the new probabilities at time $t + \delta t$. At every time interval probabilities are output to a spreadsheet, so that their behaviour over time can be monitored.

As mentioned previously, the program is initially run for 1 week (168 hours), but to make sure all probabilities get to steady-state solution, it should be run for a longer period of time. The program is therefore run for 4 weeks (672 hours) and results from the last week are shown in Figure 4.1. For illustration purposes, only a few selected probabilities are shown. As expected, over the weekend (120-168 hours), the probabilities of lower bed occupancy increase (e.g. P_8, P_{16} and P_{20}), and the probabilities of higher levels decrease (e.g. P_{24} and P_{29}).

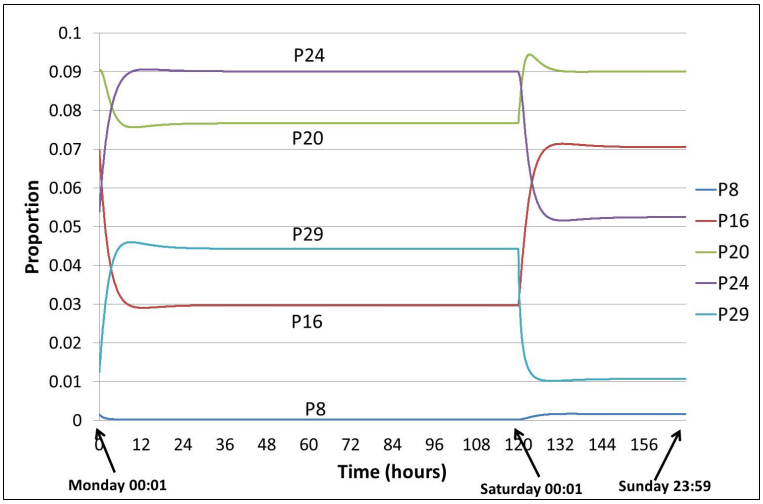


Figure 4.1: Probabilities of beds occupancies over the week

It is also of an interest how the mean and standard deviation of the bed occupancy change over the week, as shown in Figure 4.2.

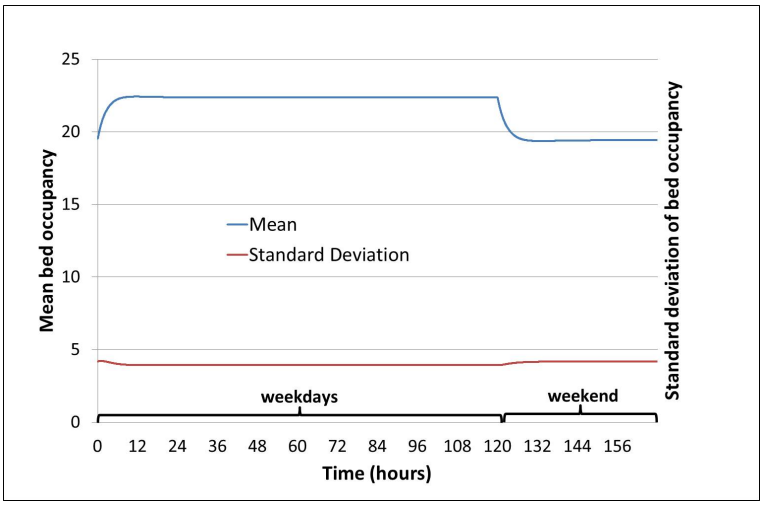


Figure 4.2: Variation in bed occupancy over the week

The blue line in Figure 4.2 indicates the mean of bed occupancy, while the red line indicates the standard deviation. As expected, the mean bed occupancy decreases on the weekend (from 22.36 to 19.40), but surprisingly, the variation increases (from 3.93 to 4.17).

The effect of increasing elective admissions on the weekend is also investigated. Very interestingly, if it is increased by 1 patient per day on the weekend, the mean and standard deviation of the bed occupancy remain almost the same throughout the whole week; it decreases from 21.52 to 22.22 and the standard deviation increases from 4.002 to 3.954.

Going one step further, and increasing elective admissions on the weekend by 1.5 patients per day would increase the mean bed occupancy by over 1 bed (from 21.52 to 22.51) and decrease the standard deviation from 4.002 to 3.91.

The transient times (times required to get from one steady-state to another steady-state level) are also obtained. If the incremented difference is less than 0.01% it is assumed that the system is at steady-state. The overall transient time was the lowest (approximately 20 hours) for bed occupancy levels between 19 and 22, which confirms what was said earlier, that the probability of having 20 beds occupied remains almost the same throughout the whole week.

The quantitative information presented is of obvious importance to the CCU Director in informing their decision making regarding bed management over the seven day cycle. The next section will investigate forecasting of future bed occupancies with probabilistic models.

4.3 Analytical Model of Bed Occupancy Predictions

A number of questions arise regarding the feasibility of 'what if' scenarios considered in Section 3.2.4 and 3.3.2. Is it possible to predict the number of beds that will be occupied in n-days time, given the bed occupancy today? Is it possible to find the most probable split between the numbers of emergency and elective patients at future days?

To accurately predict future occupancy levels based on current occupancy level, the heterogeneous nature of admissions and discharges must be taken into account. Table 4.2 gives the arrival and the discharge rates for different patient categories on each day of the week.

Table 4.2: Parameter values for admissions and discharges on each day of the week

Day of the week	Elective admissions	Emergency admissions	Elective discharges	Emergency discharges
Monday	1.1186	2.6442	0.6006	2.5399
Tuesday	1.4227	2.6795	1.0671	3.0511
Wednesday	1.6442	2.4519	1.3642	2.8785
Thursday	2.0192	2.6603	1.6656	2.8758
Friday	1.6891	2.7339	1.7508	3.0511
Saturday	0.4455	2.6124	1.5559	2.0192
Sunday	0.2628	2.6346	0.5559	1.8403

In the mathematical model the following notation is used:

B_n	bed occupancy at midnight on day n
a_n	the number of arrivals on day n
d_n	the number of departures on day n
c_n	the net increase (decrease) in bed occupancy during day n

Therefore, the number of beds occupied on day n can be expressed as:

$$\begin{aligned} B_n &= B_{n-1} + a_n - d_n \\ &= B_{n-1} + c_n \end{aligned}$$

and:

$$\begin{aligned} B_0 &= \text{bed occupancy at start of day 1} \\ B_1 &= B_0 + c_1 \\ B_2 &= B_1 + c_2 \\ &= B_0 + c_1 + c_2 \\ B_3 &= B_2 + c_3 \\ &= B_0 + c_1 + c_2 + c_3 \end{aligned}$$

In general:

$$\begin{aligned} B_n &= B_0 + c_1 + \ldots + c_n \\ &= B_0 + \textstyle\sum_{i=1}^{n} c_i \end{aligned}$$

Value of c_i can take one of the three forms:

$$c_i = \begin{cases} > 0 & \text{if } a_i > d_i, \\ 0 & \text{if } a_i = d_i, \\ < 0 & \text{if } a_i < d_i; \end{cases}$$

The probability that on day i there will be the same number of beds occupied as on day $i-1$ is equivalent to having the same number of arrivals as departures on day i, and it can be expressed as:

$$\begin{aligned} P(c_i = 0) &= P(a_i = 0) \times P(d_i = 0) + P(a_i = 1) \times P(d_i = 1) + P(a_i = 2) \times P(d_i = 2) + \ldots \\ &= \sum_{k=0}^{\infty} P(a_i = k) \times P(d_i = k) \end{aligned} \tag{4.1}$$

The probabilities of having $c_i > 0$ can be written as:

$$\begin{aligned} P(c_i = 1) &= P(a_i = 1) \times P(d_i = 0) + P(a_i = 2) \times P(d_i = 1) + P(a_i = 3) \times P(d_i = 2) + \ldots \\ &= \textstyle\sum_{k=1}^{\infty} P(a_i = k) \times P(d_i = k - 1) \\ P(c_i = 2) &= P(a_i = 2) \times P(d_i = 0) + P(a_i = 3) \times P(d_i = 1) + P(a_i = 4) \times P(d_i = 2) + \ldots \\ &= \textstyle\sum_{k=2}^{\infty} P(a_i = k) \times P(d_i = k - 2) \end{aligned}$$

$$\vdots$$

And in general:

$$\begin{aligned} P(c_i = m) &= P(a_i = m) \times P(d_i = 0) + P(a_i = m + 1) \times P(d_i = 1) + P(a_i = m + 2) \times P(d_i = 2) + \ldots \\ &= \sum_{k=m}^{\infty} P(a_i = k) \times P(d_i = k - m) \qquad \text{where } m = 1, 2, ... \end{aligned} \tag{4.2}$$

And similarly, the probability of having $c_i < 0$ can be written as:

$$\begin{aligned} P(c_i = -1) &= P(a_i = 0) \times P(d_i = 1) + P(a_i = 1) \times P(d_i = 2) + P(a_i = 2) \times P(d_i = 3) + \ldots \\ &= \textstyle\sum_{k=1}^{\infty} P(a_i = k - 1) \times P(d_i = k) \\ P(c_i = -2) &= P(a_i = 0) \times P(d_i = 2) + P(a_i = 1) \times P(d_i = 3) + P(a_i = 2) \times P(d_i = 4) + \ldots \\ &= \textstyle\sum_{k=2}^{\infty} P(a_i = k - 2) \times P(d_i = k) \end{aligned}$$

$$\vdots$$

And in general:

$$\begin{aligned} P(c_i = -r) &= P(a_i = 0) \times P(d_i = r) + P(a_i = 1) \times P(d_i = r + 1) + P(a_i = 2) \times P(d_i = r + 2) + \ldots \\ &= \sum_{k=r}^{\infty} P(a_i = k - r) \times P(d_i = k) \qquad \text{where } r = 1, 2, ... \end{aligned} \tag{4.3}$$

Since the arrivals and departures on each day of the week are assumed to be Poisson distributed, let a be the mean of the Poisson distribution for admissions and d be the mean of the Poisson distribution of the number of departures, thus:

$$P(m \text{ admissions on day } i) = \frac{a^m e^{-a}}{m!} \qquad m = 0, 1, 2, ...$$

$$P(r \text{ departures on day } i) = \frac{d^r e^{-d}}{r!} \qquad r = 0, 1, 2, ...$$

Therefore Equation 4.1 can be rewritten as:

$$\begin{aligned}
P(c_i = 0) &= \sum_{k=0}^{\infty} P(a_i = k) \times P(d_i = k) \\
&= \sum_{k=0}^{\infty} \frac{a^k e^{-a}}{k!} \times \frac{d^k e^{-d}}{k!} \\
&= \sum_{k=0}^{\infty} \frac{a^k d^k e^{-(a+d)}}{k!k!} \\
&= e^{-(a+d)} \sum_{k=0}^{\infty} \frac{(ad)^k}{k!k!} \\
&= e^{-(a+d)} \left(1 + \frac{ad}{1!1!} + \frac{(ad)^2}{2!2!} + \frac{(ad)^3}{3!3!} + \ldots\right) \\
&= e^{-(a+d)} \left(u_{0,0} + u_{0,1} + u_{0,2} + u_{0,3} + \ldots\right) \qquad (4.4)
\end{aligned}$$

where

$$u_{0,n} = \frac{(ad)^n}{n!n!}, \qquad n = 0, 1, 2, ...$$

Equation 4.2 can be rewritten as:

$$\begin{aligned}
P(c_i = m) &= \sum_{k=m}^{\infty} P(a_i = k) \times P(d_i = k - m) \\
&= \sum_{k=m}^{\infty} \frac{a^k e^{-a}}{k!} \times \frac{d^{k-m} e^{-d}}{(k-m)!} \\
&= \sum_{k=m}^{\infty} \frac{a^k d^{k-m} e^{-(a+d)}}{k!(k-m)!} \\
&= e^{-(a+d)} \sum_{k=m}^{\infty} \frac{a^k d^{k-m}}{k!(k-m)!} \\
&= e^{-(a+d)} \left(\frac{a^m}{m!} + \frac{a^{m+1} d}{(m+1)!1!} + \frac{a^{m+2} d^2}{(m+2)!2!} + \ldots\right) \\
&= a^m e^{-(a+d)} \left(\frac{1}{m!} + \frac{ad}{(m+1)!1!} + \frac{a^2 d^2}{(m+2)!2!} + \ldots\right) \\
&= a^m e^{-(a+d)} \left(u_{m,0} + u_{m,1} + u_{m,2} + \ldots\right) \qquad (4.5)
\end{aligned}$$

where:

$$u_{m,n} = \frac{(ad)^n}{(m+n)!n!}, \qquad n = 0, 1, 2, ...$$

And similarly, Equation 4.3:

$$
\begin{aligned}
P(c_i = -r) &= \sum_{k=r}^{\infty} P(a = k - r) \times P(d = k) \\
&= \sum_{k=r}^{\infty} \frac{a^{k-r} e^{-a}}{(k-r)!} \times \frac{d^k e^{-d}}{k!} \\
&= \sum_{k=r}^{\infty} \frac{a^{k-r} d^k e^{-(a+d)}}{(k-r)!k!} \\
&= e^{-(a+d)} \sum_{k=r}^{\infty} \frac{a^{k-r} d^k}{(k-r)!k!} \\
&= e^{-(a+d)} \left(\frac{d^r}{r!} + \frac{a d^{r+1}}{1!(r+1)!} + \frac{a^2 d^{r+2}}{2!(r+2)!} + \dots \right) \\
&= d^r e^{-(a+d)} \left(\frac{1}{r!} + \frac{ad}{1!(r+1)!} + \frac{a^2 d^2}{2!(r+2)!} + \dots \right) \\
&= d^r e^{-(a+d)} \left(v_{r,0} + v_{r,1} + v_{r,2} + \dots \right) \qquad (4.6)
\end{aligned}
$$

where:

$$v_{r,n} = \frac{(ad)^n}{n!(r+n)!}, \qquad n = 0, 1, 2, ...$$

Using the information shown in Table 4.2 and the obtained Formulas: 4.4, 4.5 and 4.6 it is possible to obtain the probabilities of a given change in bed occupancy given the current day of the week. It is observed that the maximum number of all admissions on one day regardless of patient type was 12, and the maximum number of discharges on one day was 13, therefore it is decided to obtain $P(c_i = s)$ where $s \in [-13, 13]$ and $s \in \mathbb{Z}$. Figure 4.3 shows how probabilities of a given bed occupancy change differs on each day of the week.

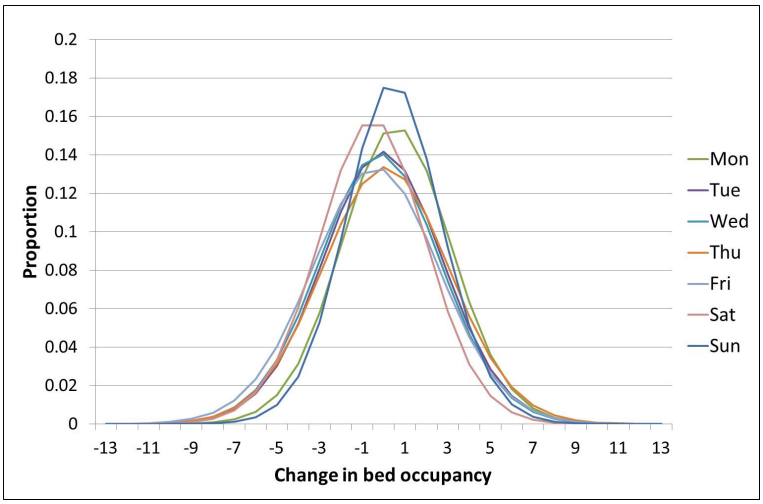


Figure 4.3: Probabilities of a given change in bed occupancy on each day of the week

The most likely bed occupancy change is 0 for most of the days with exception for Saturday, where

most likely change is -1; and for Monday, where it is 1. The expected bed occupancy change on each day of the week is also calculated using the formula: $\sum_{k=-13}^{13} kP(c_i = k)$ and, as expected, the results are significantly close to the result of the calculation: *mean number of admissions - mean number of discharges* on each day of the week and are presented in Table 4.3.

Table 4.3: Expected bed occupancy change on each day of the week

Day of the week	**Average bed occupancy change**
Monday	0.60
Tuesday	-0.05
Wednesday	-0.19
Thursday	0.09
Friday	-0.39
Saturday	-0.55
Sunday	0.48

Results from Table 4.3 show that the increase in bed occupancy is most likely to happen on Monday, Thursday and Sunday. The increase in bed occupancy on Thursday is caused by the fact that the percentage of elective admissions was highest on Thursday. Sunday and Monday had the lowest discharge rates in whole week hence increase in bed occupancy is expected. Using information obtained in this section, it is possible to obtain most likely bed occupancy three days hence, which will be shown in the following subsection.

4.3.1 Most Likely Bed Occupancy at Future Days

Having information about the probabilities of a given bed occupancy change on each day of the week, it is decided to investigate the probability of a given bed occupancy a few days in advance. As said previously, the model can produce expected bed occupancies n-days in advance; however, for the purpose of this thesis it is decided to explore bed occupancies only three days in advance. If the CCU manager wants to admit extra elective patients, when bed occupancy is low then the hospital would like to do it at a short notice, but long enough for patients who might need to make any arrangements before the planned operation.

As an illustration of the methodology, the model is demonstrated assuming a current occupancy of 23 beds. This occupancy is the most likely occupancy. Figure 4.4 illustrates how probabilities of bed occupancies change three days hence for each day of the week given today 23 beds occupied.

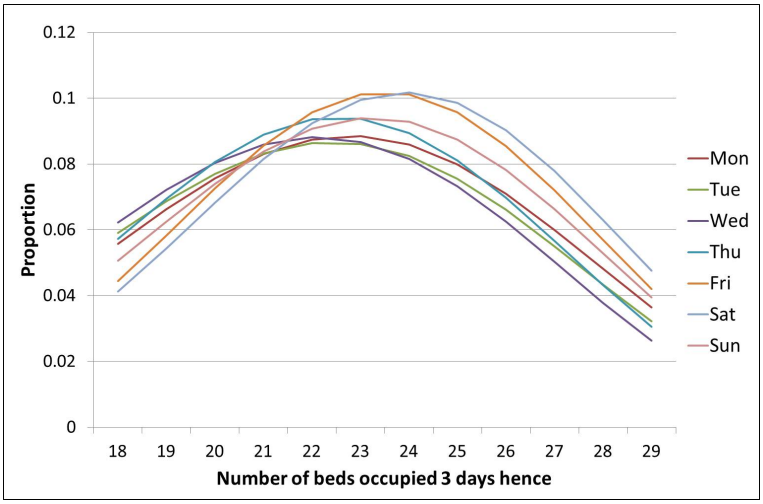


Figure 4.4: Probabilities of a given bed occupancy 3 days hence for each day of the week assuming 23 beds today

In Figure 4.4 the orange line, for example, which is marked as Friday shows how bed occupancy probabilities change three days hence. In other words, it shows Monday's bed occupancy probabilities given it is Friday today and there are 23 beds occupied today. Clearly, if the CCU Director would like to admit extra elective patients it should be done on the days when probability of high bed occupancy level is low. In this case, the lowest probability of high utilisation is for lines marked as Tuesday, Wednesday and Thursday. This means, that on those three days the decision regarding extra admission of elective patient should be made, because in three days time (Friday, Saturday and Sunday correspondingly) the bed occupancy is likely to be the lowest.

Hospital managers might be more interested in the most likely number of patients in the Unit at future days, rather than in probabilities of given bed occupancies. Figure 4.5 provides bed occupancy predictions together with the data results.

Figure 4.5 shows prediction of the most probable number of beds occupied three days hence for each day of the week. For example, if today is Monday and there are 23 beds occupied, in three days time i.e. on Thursday most likely there were 22 beds occupied according to data and model predicts 23 beds as most likely occupancy. The model gives very close predictions; the difference between the model predictions and the data is not greater than one bed for each day of the week, therefore the model can be claimed as reliable.

On Friday and Saturday there are two data bars (two shades of red) meaning that there are equally probable. Figure 4.5 also confirms what was said previously, the hospital managers should make a decisions of admitting extra elective patients on the beginning of the week, since the expected occupancy is the lowest three days hence from Tuesday and Wednesday.

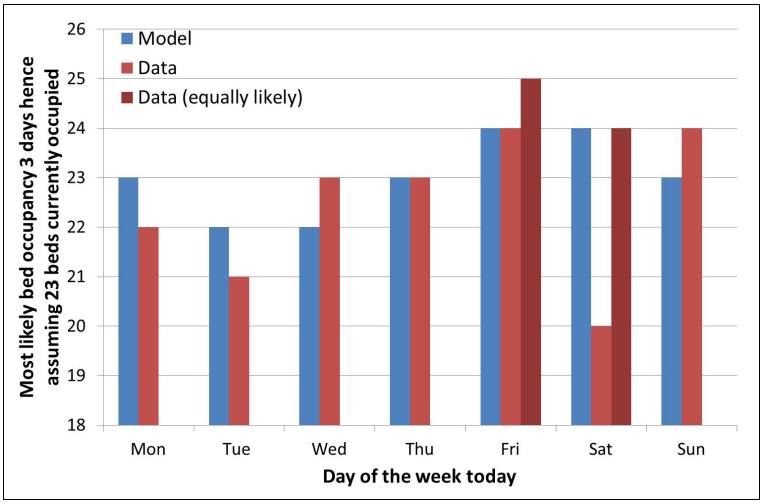


Figure 4.5: Most likely bed occupancy 3 days hence given 23 beds occupied today

The next subsection will use the same mathematical model to provide information on most likely number of emergency and elective patients at future days for each day of the week.

4.3.2 Most Probable Split Between the Numbers of Emergency and Elective Patients at Future Days

If the Director of the CCU knew today that there are, for example, 26 beds occupied, but 18 beds are occupied by elective patients an extra elective patient could be admitted in two or three days time since elective patients are very likely to stay in the CCU for a short period of time, and many of these 18 elective patients would be discharged by then. The most probable split between the numbers of emergency and elective patients at future days is explored. Again, it is dependent on day of the week today and the current split.

The notation that will be used for the current or future split is: x/y, where x denotes the number of emergency patients and y denotes the number of elective patients. Assume again that today there are 23 beds occupied in the CCU, the possible splits are: 0/23, 1/22, 2/21, ..., 23/0. The mathematical model described in Section 4.3 is utilised to find the most probable split of patients at future days. It is again decided to only show occupancy predictions three days hence, however it can easily be extended to n-days. It is also assumed that today's split is 19/4, since combination of 19 emergencies and 4 electives had the highest probability for given bed occupancy of 23. Figure 4.6 illustrate how prediction for each patient category changes for each day of the week. It shows how for each day of the week the expected number of emergency and elective patients changes for a given current split: 19 emergency and 4 elective patients. Clearly, the lowest expected number

of elective patients is 3 and is three days hence from Wednesday, Thursday and Friday. Hence, as expected, the lowest number of elective admissions falls on Saturday, Sunday and Monday. When planned admissions have the lowest expected number, the unplanned admissions have the highest.

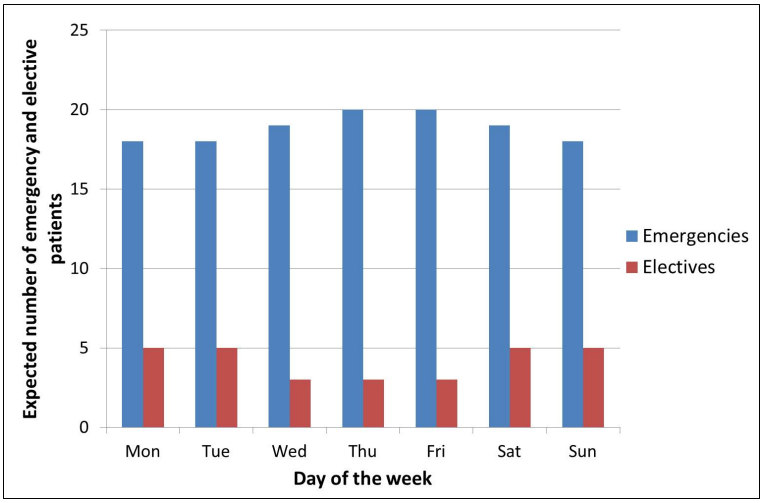


Figure 4.6: Expected number of emergency and elective patients 3 days hence given today's split is 19/4

Note that, if today is Monday and 19 beds are occupied by an emergency patients and 4 beds by an elective patients in three days time the most probable number of emergency patients is 18, the most probable number of elective patients is 5, but this does not necessarily imply that the most probable total bed occupancy will be 23.

4.3.3 Conclusions

The described mathematical model produces expected levels of bed occupancies on a 'later day' basis dependant on the day of the week today and on current bed occupancy. Information obtained in this section might be useful to the Director of the CCU to help make a decision regarding admission of extra elective patients at future days if current bed occupancy is relatively low. Also, the model provides information regarding most likely number of emergency and elective patients at future days. Different patient types require different patient to nurse ratio. Informations provided by this model can also be helpful in making decisions regarding nursing requirements on each shift. The next section will investigate the nursing levels required on each day of the week using two different approaches.

4.4 Nursing Requirements

4.4.1 Motivation of the Work

The efficient management of resources in the CCU is particularly important due to the high costs involved. The nature of the treatment administered in the CCU generates the need for costly specialist equipment and highly skilled staff. The largest proportion of the operating cost in a CCU is nursing staff (DOH 2006, [129]).

In Chapter 2 it was stated that some patients require a 1:1 nurse to patient ratio, whilst the less seriously ill require a 1:2 ratio. The Director of the CCU is faced with the difficult decision as to how many nurses to employ per shift. If too many nurses are employed, then there is a potential for large wastage in nursing costs. If too few nurses are employed, then serious consequences could arise in terms of patient care, unless the shortfall is made up by employing agency nurses. However, agency nurses can cost up to three times as much as the hospital's own nurses. As a general rule, hospital-employed nurses need to know their shift rotas significantly well into the future. The current practice is that nurses are rostered on 13-hour shifts (this allows for a 1-hour overlap with incoming nurses at the shift changeover). Two cost models to optimise the number of nurses to be employed per shift are proposed.

4.4.2 Literature Review

A literature search revealed not much previous work studies whereby staff costs are considered. Griffiths *et al.*, 2004 [75] used a mathematical model to determine the number of rostered nurses that are required to minimize overall nursing staff costs. It was concluded through use of a simulation model that nursing staff costs would have been reduced if 16 nurses have been rostered per shift rather than 14.

Harper *et al.*, 2010 [81] focused on planning the size and skill-mix of inpatient nursing teams using the capacity simulation tool, PROMPT. The approach utilises both discrete event simulation combined with a stochastic program add-on to produce optimal nursing needs by staff grade. The optimisation accounts for costs incurred for both hospital based staff and agency staff taking into account fluctuations over time of patient demands.

A growing body of research confirms the link between nurse staffing and patient outcomes. The effect of nurse staffing levels were tested against various variables, e.g. patient mortality (Aiken *et al.*, 2012 [3]), the quality of care in hospitals (Needleman *et al.*, 2002 [128]), adverse events, morbidity, medical cost (Cho, 2003 [27]), patient outcomes (Sasichay-Akkadechanun, 2003 [145] and Blegen, 1998 [16]), post-surgical adverse events (Kovner, 2002 [109]). A literature review un-

dertaken by Coombs and Lattimer, 2007 [33], focussed on organisational issues, found that staffing levels and skill mix within the CCU team can have an impact on patient outcome.

The aim of the study of Sznajder *et al.*, 2004 [154] was to propose a tool that would estimate the direct costs of stays in ICUs, which would be very useful for resource allocation inside a hospital. The model appeared to be simple and a relevant indicator which helped to organise nursing requirements.

The research of Rothberg *et al.*, 2005 [143] was conducted to determine the cost-effectiveness of various nurse staffing ratios ranging from 8:1 to 4:1. Incremental cost-effectiveness was calculated for each ratio and sensitivity and Monte Carlo analyses performed. They concluded that as a patient safety intervention, patient to nurse ratios of 4:1 are reasonably cost-effective and in the range of other commonly accepted interventions.

Section 4.4.5 contains a further literature review regarding time-dependent nursing requirements.

Finding the required number of nurses is a very complex task. For the purpose of the next Section, based on discussion with hospital managers, some realistic levels on nurse to patient ratios are chosen.

4.4.3 Nurse to Patient Ratio Required is 1:1

4.4.3.1 Model that optimises an actual expected cost

Suppose that at any particular time there are n beds occupied, and that the nurse to patient ratio required is 1:1 for all patients regardless of patients' type. The number of hospital-based nurses employed per shift is denoted by x. Let q be the cost of employing a hospital-based nurse per shift, and the cost of employing an agency-based nurse be kq. Then, the total nursing cost, $C(x, n)$, may be expressed as:

$$C(x,n) = \begin{cases} qx, & \text{if } x \geq n \\ qx + (n-x)kq, & \text{if } x < n \end{cases}$$

and the expected cost, where P_n is the probability of having n patients present in the system and c is the system capacity, is therefore given by:

$$\begin{aligned} E(cost|x) &= \sum_{n=0}^{c} P_n C(x,n) \\ &= \sum_{n=0}^{x} P_n qx + \sum_{n=x+1}^{c} P_n \left(qx + (n-x)kq\right) \end{aligned} \tag{4.7}$$

The probabilities of having n beds occupied are taken from the data (but analytical expressions can also be used) and nursing costs are obtained for appropriate values of x, where $x = 10, 11, ..., 29$. It is assumed that $q = 1$ (cost of hospital-based nurse) and $k = 3$ (cost of employing an agency-based nurse is three times more expensive than the hospital-based nurse). Figure 4.7 displays the expected nursing cost for different number of hospital-based nurses. The optimal number of hospital-based nurses that should be employed per shift is 23.

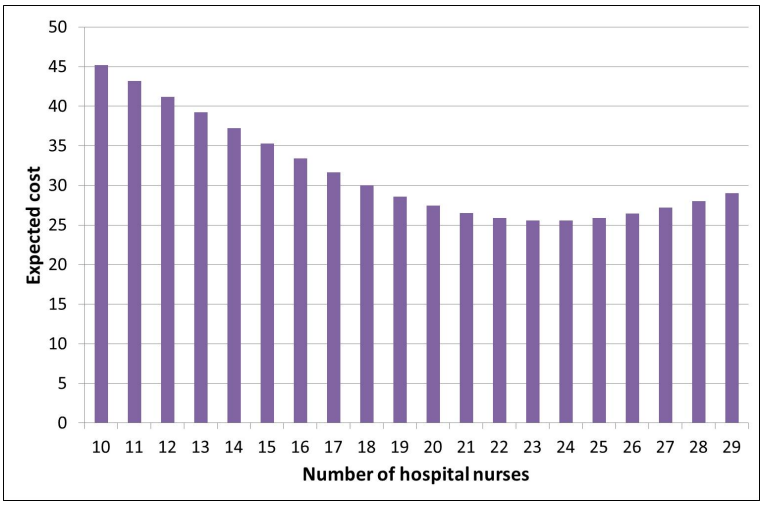


Figure 4.7: The expected nursing cost for different number of hospital nurses

4.4.3.2 Newsboy Model

It is worth noting that this problem of nurse staffing can also be approached using a stochastic inventory model. Indeed, the Newsboy Model (Winston, 1998 [168]) models a stock control problem to which is associated C_o, the cost of having too many hospital-based nurses (oversupply cost) and C_u, the cost of having not enough hospital-based nurses (undersupply cost). Let $F(x)$ be the cumulative distribution function (CDF) of the number of beds occupied. Then, by the Newsboy Model the expected cost is minimised by the x which satisfies:

$$F(x-1) < \frac{C_u}{C_u + C_o} \leq F(x)$$

The problem can be modelled with $C_u = k - 1$ and $C_o = 1$, in this case $k = 3$. Therefore $\frac{C_u}{C_u+C_o} = \frac{2}{3}$. Table 4.4 shows the CDF and it can be seen that the Newsboy Model approach to this problem also recommends 23 nurses.

Table 4.4: Cumulative distribution function representing the demand for nurses

x	$F(x)$
20	0.361
21	0.461
22	0.559
23	**0.671**
24	0.770
25	0.850
26	0.912

4.4.4 The Nurse to Patient Ratio is Variable

4.4.4.1 Model that Optimises an Actual Expected Cost

When a patient is admitted, a decision is made on the basis of the information available as to the level of care required. A Level 3 patient requires a 1:1 ratio, while Levels 1 and 2 require 1:2. However, patients' conditions may deteriorate or improve during their stay in the CCU, and hence the required level of nursing may also change. A reasonable proxy is to assume that elective patients require Level 1 or 2 care (i.e. one nurse is required to two elective patients), and emergency patients require Level 3 care (i.e. one nurse to one emergency patient). Therefore, the probabilities of having i emergency patients and j elective patients, $P_{i,j}$, as determined by Equation 3.12, needs to be employed. The number of nurses required is then given by $i + \left\lceil \frac{j}{2} \right\rceil$ i.e. one nurse to an emergency patient and a half nurse to an elective patient. The number of required nurses, r, is given by:

$$r = r(i,j) = i + \left\lceil \frac{j}{2} \right\rceil = \begin{cases} i + \frac{j}{2}, & \text{if } j \text{ even} \\ i + \frac{j+1}{2}, & \text{if } j \text{ odd} \end{cases}$$

Consequently, the total nursing cost is:

$$C(x;i,j) = \begin{cases} qx, & \text{if } x \geq i + \left\lceil \frac{j}{2} \right\rceil \\ qx + (i + \left\lceil \frac{j}{2} \right\rceil - x)kq, & \text{if } x < i + \left\lceil \frac{j}{2} \right\rceil \end{cases}$$

Therefore, the probability that exactly r nurses are needed is given by:

$$P(r) = P_{r,0} + \sum_{i=0}^{r-1} \left(P_{i,2(r-i)} + P_{i,2(r-i)-1} \right)$$

and so the CDF for the nurse demand is given by:

$$F(r) = P(X \leq r) = \sum_{m=0}^{r} \left(P_{m,o} + \sum_{i=0}^{m-1} \left(P_{i,2(m-i)} + P_{i,2(m-i)-1} \right) \right) \tag{4.8}$$

The expected cost of employing x nurses on each shift is now given by:

$$\begin{aligned} E(C(x;i,j)) &= \sum_{i=0}^{c} \sum_{j=0}^{c-i} P_{i,j} C(x;i,j) \\ &= \sum_{r=1}^{c} P_{r,0} C(x;i,j) + \left(\sum_{i=0}^{r-1} P_{i,2(r-i)} C(x;i,2(r-i)) + P_{i,2(r-i)-1} C(x;i,2(r-i)-1) \right) \\ &= \sum_{r=1}^{x} \left(P_{r,0} qx + \sum_{i=0}^{r-1} \left(P_{i,2(r-i)} qx + P_{i,2(r-i)-1} qx \right) \right) + \\ &+ \sum_{r=x+1}^{c} \left(P_{r,0} \left(qx + (r-x)kq \right) + \sum_{i=0}^{r-1} \left(P_{i,2(r-i)} \left(qx + (r-x)kq \right) + P_{i,2(r-i)-1} \left(qx + (r-x)kq \right) \right) \right) \end{aligned} \tag{4.9}$$

Using historical data, the probabilities of having $n = i+j$ beds occupied are obtained and therefore probabilities of having i emergency and j elective beds, $P_{i,j}$, occupied are found using Formula 3.12. Again, it is assumed that $q = 1$ and $k = 3$. Figure 4.8 displays the expected nursing cost for different number of hospital-based nurses.

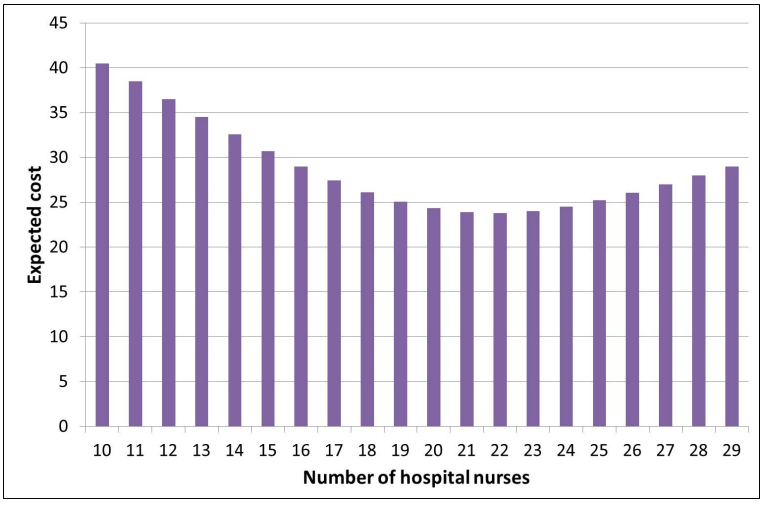


Figure 4.8: The expected nursing cost for different number of hospital nurses

It can be seen, that the optimal number of the hospital-based nurses that should be employed per shift is now 22, which gives the minimum expected cost of 23.80. As expected, the optimal number of nurses is lower than previously when the nurse to patient ratio was 1:1 irrespective of patient type.

4.4.4.2 Newsboy Model

The Newsboy Model can be used to solve the nurse staffing problem when the number of required nurses depends on the number of emergency, i and elective, j patients. Recall, the expected cost is minimised by the x which satisfies:

$$F(x-1) < \frac{C_u}{C_u + C_o} \leq F(x)$$

The associated costs, C_o and C_u have not changed, so the ratio is still $\frac{C_u}{C_u+C_o} = \frac{2}{3}$.
Also recall Formula 4.8 that the CDF is as follows:

$$F(x) = P(\leq x) = \sum_{m=0}^{x} \left(P_{m,0} + \sum_{i=0}^{m-1} P_{i,2(m-i)} + P_{i,2(m-i)-1} \right)$$

Table 4.5 shows values for the CDF and it can be seen that the Newsboy Model approach to this problem also recommends 22 nurses.

Table 4.5: Cumulative Distribution Function representing the demand for nurses

x	$F(x)$
19	0.417
20	0.525
21	0.637
22	**0.743**
23	0.833
24	0.900
25	0.948

The difference in these two approaches is that the first optimises an actual expected cost, whereas the Newsboy model optimises the expected wastage costs.

4.4.5 Time Dependent Nursing Requirements

The scheduling of nurses is not a trivial task. This is due to the large fluctuations in numbers required during the quiet and busy periods. It is assumed that a fixed number of nurses is scheduled per shift, regardless of the actual number required. Recall that bed occupancy, as well as the number of elective admissions depends on the day of the week. Intuitively, the number of nurses required on the weekend should be lower than during the week. This section will investigate how nursing requirements change during the week.

Whitt, 2007 [166] produced a comprehensive review of time-dependent staffing requirements literature. Setting staffing requirements of a service system with time varying demand is a major challenge. It has attracted a significant body of research over the last three decades, and a few approaches have been developed as a result; Vassilacopoulos, 1985 [160] used a deterministic model for allocation physicians to weekly shifts in an A&E department. They set physician levels proportional to the hourly mean arrival rates. Coats and Michalis, 2001 [30] used simulation to compare two different shift patterns with the existing one in A&E. Green *et al.*, 2006 [71] modelled a local emergency department in the US as a single station queueing system to determine physicians staffing. Vile, 2012 [163] developed time-dependent priority queueing system to determine the optimal staffing levels for the WAST (Welsh Ambulance Service Trust). Whilst most of the research has concentrated on single service systems, Izady and Worthington, 2012 [93] proposed staffing algorithm which relies on infinite server networks to compute the resources' time dependent workloads. The authors showed how queueing models equipped with simulation can be used to alleviate the congestion problem of emergency departments by modifying the staffing profiles. Agnihothri and Taylor, 1991 [2] sought the optimal staffing at a hospital scheduling department that handled phone calls whose intensity varies throughout the day. The paper grouped periods that receive similar call intensity and determines the necessary staffing for each such intensity, so that staffing varies dynamically with call intensity.

The time-dependent optimal nursing requirement is found using Euler's method. The program that was written in Visual Basic to solve the time-dependent equations is adjusted so that at every time-step $\delta t = 0.05$ (hours) it calculates the number of required nurses by optimising the cost given by Formula 4.9, which depends on the number of emergency and elective patients present at time δt. As said previously, the number of elective admissions is dependent on the day of the week. For each day of the week different value of λ_2 (number of elective admissions) is used. The parameter values are presented in Table 4.6. The program is run for 4 weeks in order to get to steady-state solution and results from the last week are displayed in Figure 4.9.

As expected, the number of nurses that are required to work during weekends is 21 and is lower than during the weekdays, when 22 nurses are required. Throughout weekdays no variation is observed in the number of nurses required.

Table 4.6: Number of elective admissions dependent on day of the week

Day of the week	**Value of λ_2**
Monday	1.1118
Tuesday	1.4249
Wednesday	1.6326
Thursday	2
Friday	1.6837
Saturday	0.4441
Sunday	0.2619

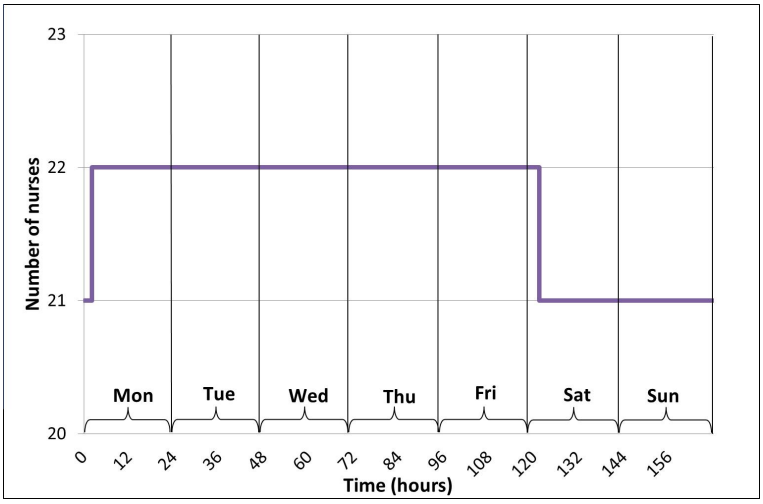


Figure 4.9: Time dependent number of required hospital-based nurses

4.5 Conclusions

The objective of this chapter has been to consider many aspects of theoretical and practical application of mathematical modelling in the CCU setting. This chapter considered time-dependent aspects of bed probabilities. The presented information regarding prediction of future bed occupation is of great importance to the Director of the CCU. Also, very importantly from a cost point of view the problem of how many nurses should be employed per shift was addressed in this chapter. Two cost models to optimise the number of nurses have been proposed. Both models recommended the same number of nurses required on each shift.

Chapter 5

Data Analysis of Two Data Sets From the Royal Gwent and the Nevill Hall Critical Care Units

5.1 Introduction and Motivation of the Study

The project described in the next two chapters was initiated by managers from the Aneurin Bevan Local Health Board. The initial proposition of the project was to investigate the effect of bed blocking and patients' transfers between the Royal Gwent and the Nevill Hall hospitals using queueing theory and similar methods to those described in Chapters 2 and 3. The study described in Section 6.5 was a consequence of an ongoing NHS project of building a new hospital, that will replace both existing hospitals.

This study describes the project undertaken with the managers from the Aneurin Bevan Local Health Board, which is an NHS Wales organization in South Wales, that serves 21% of the total Welsh population ([89]). Critical care is delivered on two sites, at the Nevill Hall hospital in Abergavenny and at the Royal Gwent hospital in Newport. For the remaining of this thesis, the Royal Gwent hospital will be referred to as RG and the Nevill Hall hospital as NH.

The data used for analysis was provided by the Intensive Care National Audit and Research Centre (ICNARC) and refers to patients admitted to CCUs in RG and NH, and includes the period of three years, from the 1^{st} January 2009 till the 31^{st} December 2011. The ICNARC system was established by clinicians and it collects detailed information; about half of all CCUs across England and Wales currently use this system. The data set contains information about a patient's source of admission, date and time of admission, date and time of discharge, CCU outcome and delay to discharge. In NH, the smaller of the two hospitals, there were 8 beds available throughout the analysed period. In

RG the bed capacity changed twice. In the beginning of the analysed period the Unit had 14 beds. In September 2010, after 20 months, the bed capacity was increased by one, to 15, and in October 2011 it was increased by one, giving a total bed capacity of 16.

When planning CCU capacity, most units have to take into account emergency and elective admissions; however, as informed by the hospital managers the percentage of planned admissions in NH and RG is very low (below 5%), hence the split will not be taken into account in this case.

The initial objective of this chapter is to analyse the actual data from both CCUs to determine admission and service patterns, delay to discharge and flow of patients through the Units. Any analysis described in this chapter will account separately for RG and NH. The primary objective of this study is to develop a valid mathematical model of the CCUs, which can be used as a tool for management.

5.2 Data Analysis

The aim of this part of the project is to understand the current system of CCUs, and to expand on this by modelling new scenarios and forecasting the outcome. By analysing the data provided by the CCUs, it is possible to establish the admission rates and service rates of patients, which characterise the flow of patients through the Units.

The NH data includes information regarding 1640 patients and the RG data regarding 2458 patients. NH CCU receives patients from 27 different sources, but the majority of patients come from Recovery / Theatre (30%), A&E (24%) and Emergency and Assessment Unit (11%). Patients from 76 different sources are admitted to RG CCU, with a majority from Recovery / Theatre (31%), A&E (26%) and Medical Assessment Unit (5%). Information was only available for patients that were admitted to the Unit, information regarding patients rejected due to insufficient number of beds was not available.

This section will consider four different aspects for both CCUs and the structure is as illustrated in Figure 5.1.

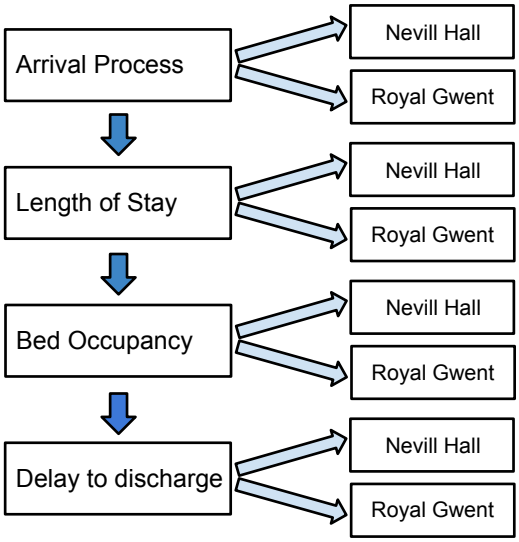


Figure 5.1: Structure of the data analysis section

5.2.1 Admission process

This section describes analysis of the data, so that the nature of admissions to the CCUs can be explored in greater depth. The main objective is to determine the appropriate statistical distributions, that would represent the profile of admission to the CCUs in both hospitals in the queueing model. Admission processes to each CCU will be considered separately.

5.2.1.1 Nevill Hall

The hourly admission pattern is represented graphically in Figure 5.2. It displays the proportion of admissions according to various hour of the day.

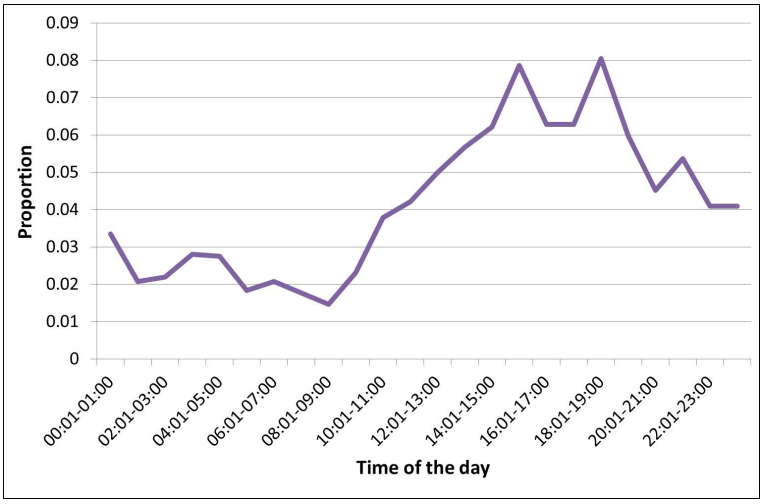


Figure 5.2: Proportion of admissions at each hour of the day

There is a visible hourly arrival pattern; there are very few admissions in the morning hours with mainstream of admissions in the afternoon or evening, with highest peaks between 3pm and 4pm and between 6pm and 7pm. Most surgeries are scheduled to start in the morning and correspond to admissions at the CCU in the afternoon, hence the peaks.

Figure 5.3 displays the daily admission patterns. It shows that the admission process does not differ very much from Monday to Friday; the weekday average is 1.62 patients per day. On the weekend, however, there are fewer admissions, 1.18 on average, suggesting that the clinicians prefer not to work at the weekends.

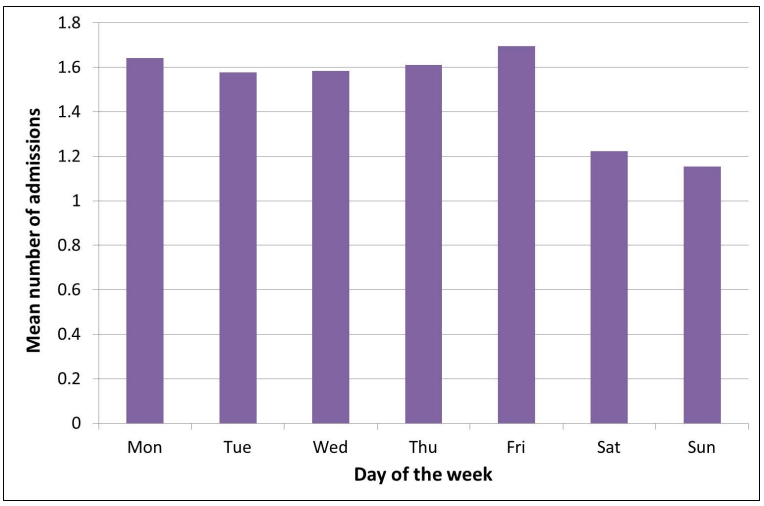


Figure 5.3: Mean number of admissions on each day of the week

The monthly average admissions are also considered, and the rates range from 1.35 (in August) to 1.63 (in February) patients per day, which are most likely related to the weather conditions. Harper *et al.*, 2012 [80] showed that accurately forecasting of number of hospital admissions is significantly improved by incorporating meteorological information.

Summary statistics for the daily number of arrivals at the CCU in NH are given in Table 5.1.

Table 5.1: Summary Statistics for the number of admissions on each day

Summary statistic	**Value**
Mean	1.4977
Median	1
Standard Deviation	1.1653
Minimum	0
Maximum	6

The frequencies and similar values for the mean and variance indicate that the admission process could be modelled as a Poisson process. The next task is to determine a value for the parameter λ which would provide the appropriate fit to the distribution of daily admissions. This is achieved using the Excel plug-in: *@Risk*, a distribution fitting software. A value of $\lambda = 1.5245$ is found, which gives a very low value (0.001187) for the sum of the squares of the deviations. Figure 5.4 displays a frequency distribution of the number of admissions on each day along with the fitted Poisson distribution, confirming goodness of fit.

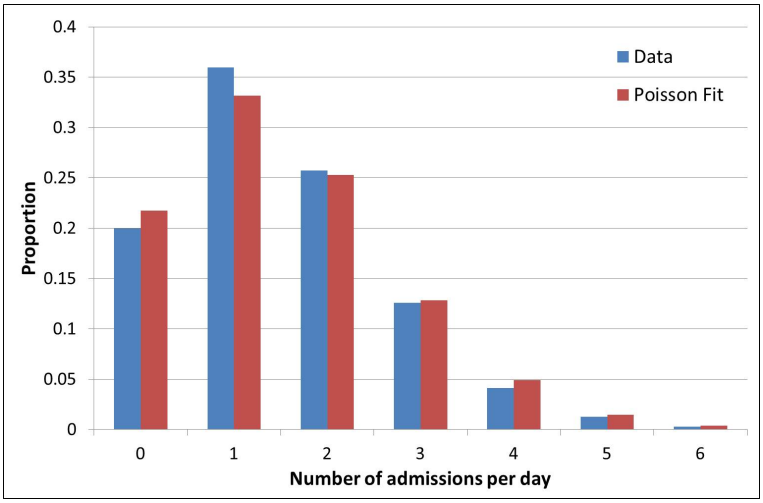


Figure 5.4: Poisson fit to the distribution of admissions

The pattern of discharges by time of day and day of the week are also examined. Patients are mostly discharged in the afternoon. This is explained by the fact that staff would have had enough time to assess patients' conditions, and ensure there is available bed in an ordinary ward, where patients will be transferred. Only 8% of the total discharges occur between midnight and 9am, and a majority of those discharges happened as a result of death. It appears that discharges' hourly peaks are shifted to the left of admissions' hourly peaks. This is because the CCU managers want to accommodate beds for the coming surgery patients. Profile of weekly discharges shows that the average number of discharges is steady throughout the weekdays, with average of 1.64, and reaches its maximum on Thursday (1.77). The mean number of discharges decreases to 1.13 on the weekends confirming again the fact that there is fewer qualified staff working on the weekends.

5.2.1.2 Royal Gwent

Similar consideration regarding admissions to the CCU at RG will now be undertaken. Recall that the bed capacity in RG was changed twice, therefore each of the three periods will be considered separately.

The hourly admission pattern was initially considered for each period. However, the pattern was found to be very similar across the three periods, hence the overall hourly admission pattern is considered for the whole study period. Figure 5.5 displays the proportion of admissions according to various hour of the day.

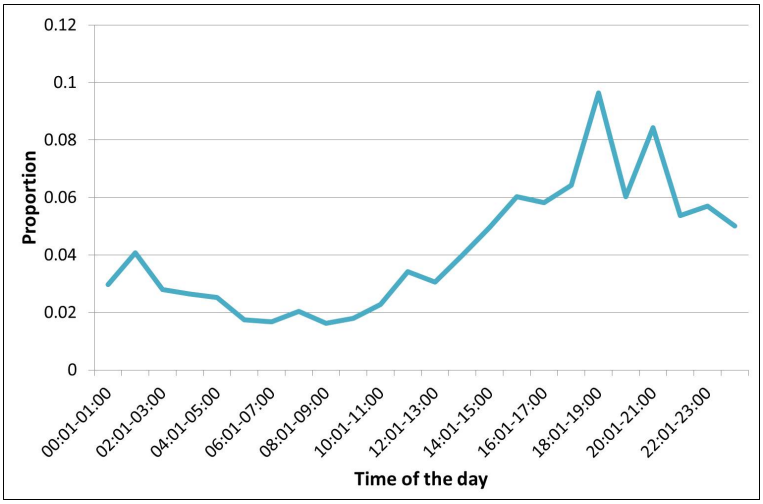


Figure 5.5: Hourly proportion of admission times

There is a visible pattern in the time of admission, as for NH. Very few admissions during the morning hours and a majority of admissions during evening hours. The highest peaks of admissions are between 6pm and 7pm and between 8pm and 9pm, so both peaks are slightly shifted to the right compared with NH times of admissions.

Any daily admission trends are also examined; Figure 5.6 illustrates them.

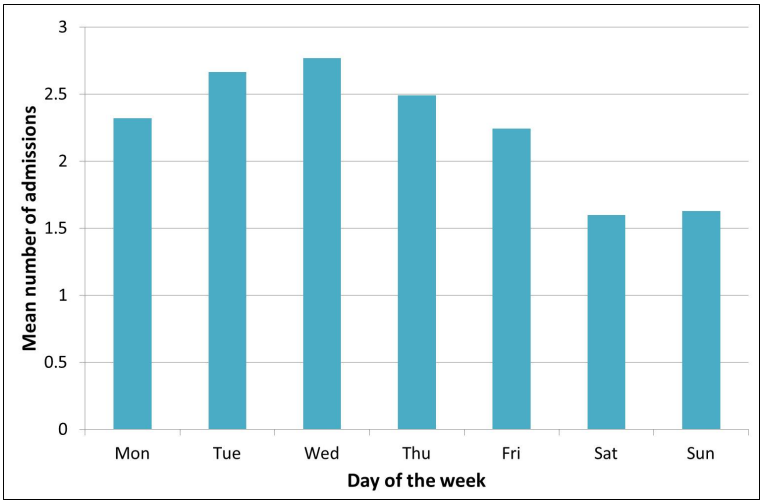


Figure 5.6: Daily average number of admissions

Figure 5.6 shows characteristic pattern of admissions: the average number of admissions increases in the beginning of the week and starts to decrease towards the weekend. The average number of admissions on the weekdays and weekends differs by nearly one patient per day; on average 2.50 patients per day on weekdays compared with 1.61 on weekends.

The monthly admissions shows that on average there are more admissions in the winter months (September-December). The rates ranged from 2.11 (in May) to 2.42 (in September), again suggesting that admissions are dependent on the weather conditions. As mentioned previously, this has been investigated previously by Haper *et al.*, 2012 [80].

The admission process needs to be considered for modelling purposes. Summary statistics for the daily admission numbers at the CCU in RG, separately for each of the three periods and overall are given in Table 5.2.

Table 5.2: Summary Statistics for the number of admissions on each day

Summary Statistic	**Period 1**	**Period 2**	**Period 3**	**Overall**
Mean	2.3026	2.1722	2.1739	2.2447
Median	2	2	2	2
Standard Deviation	1.4523	1.3763	1.298	1.4131
Minimum	0	0	0	0
Maximum	8	7	6	8

The overall average number of admissions is 2.24 and there is a little variation across the three periods. Intuitively, having more beds available would increase the number of admissions; however, it has not been observed. In the first period, which lasted 52% of the time, the daily admission rate is 2.30 and it is marginally decreased to 2.17 in Period 2 and remains very close in Period 3.

As the mean number of admissions in the three periods is similar, it is decided that the combined overall frequencies will be considered. An inspection of the frequencies and similar values for the mean and variance suggest that the number of daily admissions may be modelled by a Poisson distribution. The found parameter $\lambda = 2.2556$ provides the most appropriate fit to the distribution of daily admissions, giving a very low value (0.00086) for the sum of the squares of the deviations. Figure 5.7 demonstrates a frequency distribution along with the fitted Poisson distribution.

The pattern of discharges by time of day and day of the week are also analysed. Patients are mostly discharged in the afternoon and evening. Between midnight and 9am there were only 8.6% of the total discharges and it is found that the majority of those discharges happened as a result of

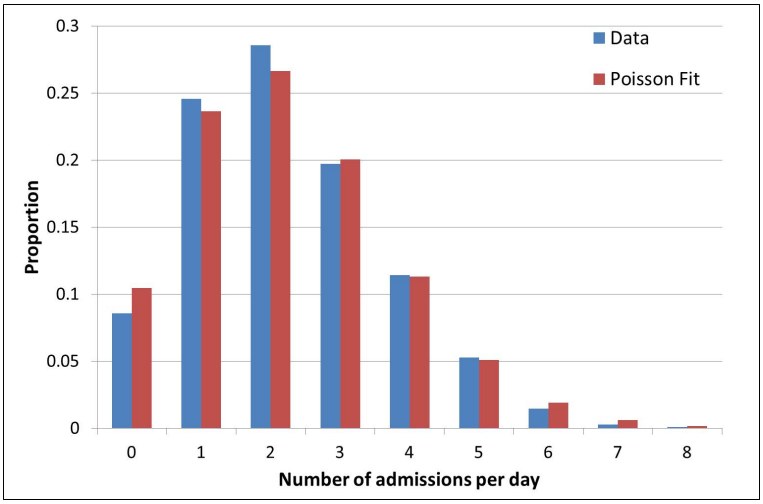


Figure 5.7: Poisson fit to the distribution of admissions

death. There is first a peak of discharges between 4pm and 5pm, followed by peak of admissions between 6pm and 7pm. Attention is taken to the profile of weekly discharges, the average number of discharges steadily increases from Monday to Friday to reach its maximum (2.82 discharges) on Friday, and there is a big reduction in discharges at the weekends (1.5 discharges). The peak of discharges on Friday suggests that the staff prepare the Unit in case of a big influx of patients on the weekend.

Finally, it is checked whether admission day has an impact on day of discharge. It is concluded that for all days except for Saturday, the most probable day of discharge is the next day of the week; for example if a patient is admitted on Monday, their most likely discharge day will be Tuesday. The exception is Saturday, when a patient is equally likely to be discharged on Sunday or Monday.

5.2.2 Length of Stay

Patient's LoS varies widely; some patients are discharged within 1-2 days, after post operative observation, others require life support machines for several weeks or even months. The main objective of this section is to determine distributions for each hospital that may accurately generate the length of stay (LoS) in the CCU.

The LoS can be influenced by the CCU bed occupancy level. This fact is confirmed by the hospital managers, who claim that some of their decisions are dependant on the number of occupied beds. This finding was also highlighted by Mallor, 2011 [117].

The data sets provide information for each admitted patient regarding date and time of admission and date and time of discharge. LoS is calculated for each patient, and the LoS profile for patients

in each hospital will be considered separately.

5.2.2.1 Nevill Hall

A number of relevant summary statistics of the data are considered and are presented in Table 5.3.

Table 5.3: Summary statistics for LoS in NH

Summary statistic	**Value (days)**
Mean	3.8245
Median	1.7100
Standard Deviation	6.5326
Minimum	0
Maximum	86.7535

Clearly, the data contains a vast range of values as indicated by the standard deviation of 6.53 days; on average patients stayed in the CCU for 3.82 days; however, some patients required life support for as long as 86.75 days.

After inspection of the mean and standard deviation for the overall LoS the initial suggestion is that a Negative Exponential distribution would not provide a reasonable fit, since the mean is not similar to the standard deviation. However, the lowest value of the sum of the square of the deviation (0.002) is achieved for the Negative Exponential distribution with the parameter $\mu = 0.4096$, which is accepted as a good fit. Consider Figure 5.8, which presents the distribution of the LoS in the CCU against the fitted Negative Exponential distribution. Each bar represents a one day period; 2.98% of observations (LoS greater than 20 days) are excluded from the graph for presentation purposes.

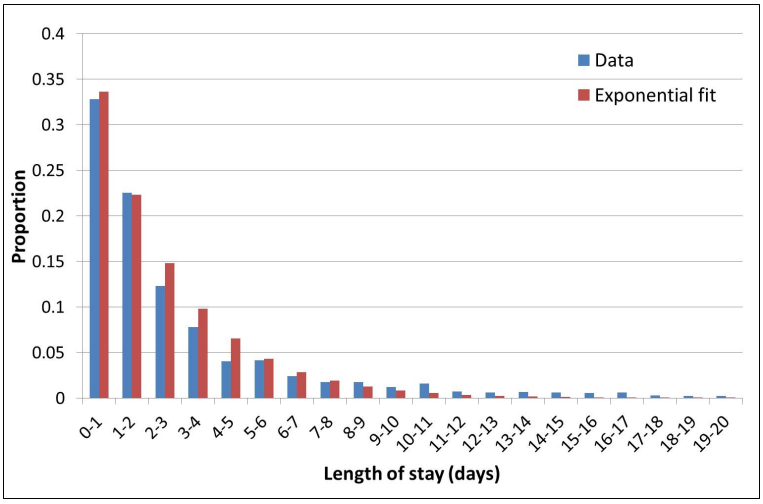


Figure 5.8: Length of stay distribution with Negative Exponential fit

The average LoS for different age groups is considered; patients in the age group between 61 and 70 years, which accounts for 21% of total admissions, have the longest average LoS (4.68 days). The shortest average LoS (0.23 days) is observed for patients in the age group between 0 and 10 years, however, that group accounts for only 3% of the total admissions.

In the first part of this thesis, consideration was given to the average LoS for emergency and elective patients. It was shown that elective patients needed support for significantly shorter period of time. In the current project the average LoS is found for patients admitted from the most common three sources of admission: Recovery / Theatre, A&E and Emergency Assessment Unit (EAU) which in total account for 64% of all admissions. Table 5.4 shows the three different admission sources with their average LoS for each source in an increasing LoS order.

Table 5.4: Average length of stay in NH for different admission sources

Admission source	**Probability of admission from given source**	**Average LoS (days)**
Recovery / Theatre	30%	2.3645
A&E	24%	3.1960
EAU	10%	4.7109
other	36%	5.1908

The recovery process is dependent on admission source; patients admitted following surgery are expected to make a quick recovery. The mean LoS of patients in this category is below average (2.36 days compared with the overall average of 3.82 days). Patients are referred to the EAU by a doctor in the Emergency Department for a specialist medical opinion on their condition; mean LoS in this category is above overall average, 4.71 days.

There is no clear reason why patients who are admitted on one day of the week should stay in CCU longer than those admitted on any other. Having said that, patients who are admitted on Sunday have the longest average stay in hospital, while patients admitted on Wednesday usually have the shortest stay (4.76 days on Sunday compared with 3.03 days on Wednesday). This could be explained by the fact that Sunday has the smallest probability of admission, and probably only the most ill patients who require longer life support are admitted to the CCU. The found results are the reverse of results found in the Bed Management Audit Commission Report, 2003 [7], where patients admitted on Thursday had the longest stay in CCU while patients admitted on Sunday usually had the shortest stay.

Next, consideration will be taken to patients' LoS admitted to RG.

5.2.2.2 Royal Gwent

Recall, the data set was divided into three periods, summary statistics of LoS (in days) of patients admitted in each of the three periods to RG CCU are presented in Table 5.5.

Table 5.5: Summary statistics for the length of stay in RG (days)

Summary statistic	**Period 1**	**Period 2**	**Period 3**	**Overall**
Mean	4.5264	5.7989	5.4346	5.0447
Median	2.3854	3.1	3.3	2.8
Standard Deviation	8.9936	9.3329	7.8261	9.0388
Minimum	0.0208	0	0	0
Maximum	193.0938	152.4	61	193.0937

The data shows high variation across the three periods. Intuitively, having more beds should not increase patient's duration of stay, however this is observed. The average LoS in Period 1 is 4.53 days and it is increased by 1.27 days in Period 2. In Period 3, the average is marginally decreased by 0.37 days to 5.43 days. The data contains a very wide range of values; on average patients stayed in the CCU for 5.04 days; however, there were patients who required life support for over six months.

After inspection of the mean and standard deviation for the overall LoS the initial suggestion is that a Negative Exponential distribution would not provide a reasonable fit, since the mean is not similar to the standard deviation. However, the lowest value of the sum of the square of the deviation (0.00127) is achieved for the Negative Exponential distribution with the parameter $\mu = 0.26087$, which is accepted as a good fit. Consider Figure 5.9 which presents the distribution of the LoS in the CCU against the fitted Negative Exponential distribution. 1.01% of observations (LoS greater than 40 days) are excluded from the graph for presentation purposes.

The average LoS for different age groups is considered; patients in the age group between 51 and 60 years, which accounts for 15% of the total admissions, have the longest average LoS (6.21 days). The shortest average LoS (2.5 days) is observed for patients in the age group between 11 and 20 years; however, that group accounts for only 2.4% of the total admissions. Elderly patients are more likely to stay longer as inpatients compared to younger ones. They are more prone to hospitalisation related complications such as infections or worsening of their condition.

It is also checked whether the source of admission has an effect on LoS. The average LoS is found for patients admitted from the three most common sources of admission, namely: Recovery/Theatre, A&E and MAU, which in total account for 62% of all admissions. Table 5.6 shows the

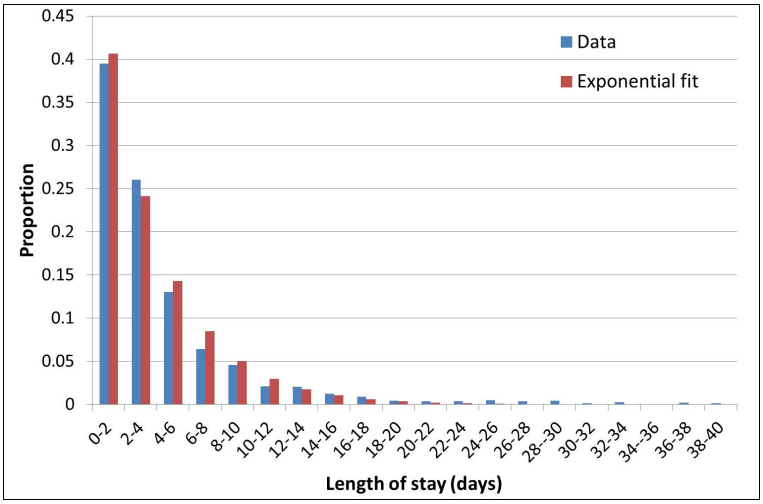


Figure 5.9: Length of stay distribution with Negative Exponential fit

three different admission sources and the average LoS for each source in an increasing LoS order.

Table 5.6: Average length of stay in RG for different admission sources

Admission source	**Probability of admission from given source**	**Average LoS (days)**
Recovery / Theatre	31%	4.2801
A&E	26%	4.3183
MAU	5%	5.5087
others	38%	6.1004

Visibly, recovery process in RG is dependent on the admission source; patients admitted for post operative monitoring are expected to stay in the CCU the shortest time, 4.28 days on average. Patients admitted from A&E are also expected to make a fairly quick recovery. The mean LoS of patients in this category is below average (4.31 days compared with the overall average of 5.04 days). Patients admitted from other, less likely, sources are expected to stay in the CCU for longest (6.10 days), on average one day longer than the overall average LoS.

Finally, it is checked whether day of the week when patients are admitted has an impact on their LoS. Patients admitted on Thursday have on average longest LoS, which confirms findings quoted in the Bed Management Audit Commission Report, 2003 [7]. The shortest average LoS is observed for patients admitted on Monday, though it is unlikely that their cases are less complex. There is no justifiable reason why that happens.

5.2.3 Bed Occupancy

Bed occupancy measure is not provided directly by the data set, but can be obtained by examining the admission and discharge dates and times. A program written in VBA read in the number of patients present in the Unit every hour from 01/04/2009 to 31/12/2011 and the bed occupancy is output to a worksheet. The reason for omitting a period of three months is to avoid underestimation; patients could have been admitted for example on 31/12/2008 and stayed in the Unit for a month. Records of that patient would not be included in the data set and therefore the fact that the patient occupied a bed would have been skipped. Occupancy census for each CCU will be measured separately.

5.2.3.1 Nevill Hall

Nevill Hall is the smaller of the two considered hospitals, and the CCU consists of 8 beds. The number of beds occupied at every hour in study period is illustrated in Figure 5.10.

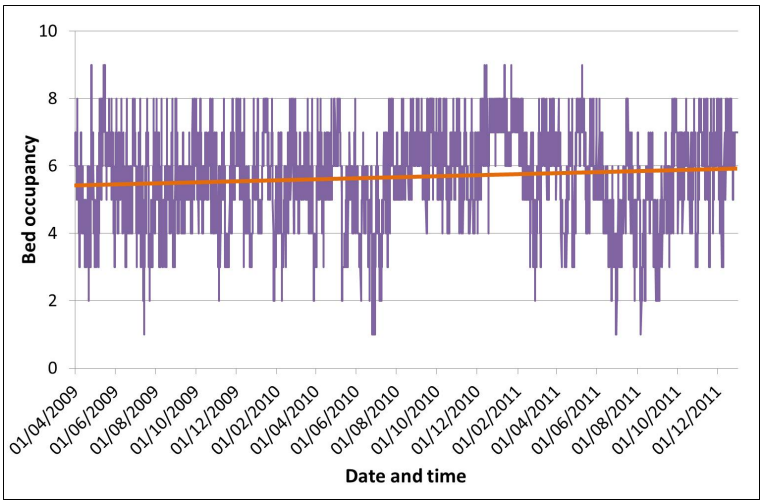


Figure 5.10: Hourly bed occupancy at NH (April 2009-December 2011)

The average number of beds occupied in the study period is 5.67 with standard deviation of 1.47. Very often hospital managers use the term utilisation rate instead of the average number of beds occupied. Utilisation rate is the ratio of mean bed occupancy to the bed capacity. In NH, the average utilisation rate is 70.9%, which is considerably low, since the most quoted bed occupancy target in the literature is 85% ([64]). The number of beds occupied fluctuates up to 9 during busy periods and dropping to as low as 1 at quiet times. An initial analysis reveals that on 0.3% of occasions there are more than 8 patients in the Unit, being probably treated in the Recovery Room after their surgery until a CCU bed becomes available. The orange increasing trend-line in Figure 5.10 suggests that the bed utilisation rates are increasing with time.

The data is investigated to check for any time dependencies; whether on particular days of the week it is more likely that occupancy will be lower or higher than on others. The weekly trends of bed occupancy are investigated and there are no visible trends. The average bed occupancy fluctuates between 5.60 (on Saturday) and 5.72 (on Sunday). Also, hourly trends are examined; it appears that on average a day starts with the lowest average bed occupancy (5.55 beds), it then steadily increases to reach its maximum at 2pm (5.80 beds) and then starts to decrease again towards the evening.

Consider the frequency distribution of the bed occupancy in the CCU during the study period, presented in Figure 5.11. Clearly, the most likely bed occupancy is 6, with corresponding probability of 26.6%. The Unit was full to its capacity on 9.6% of occasions.

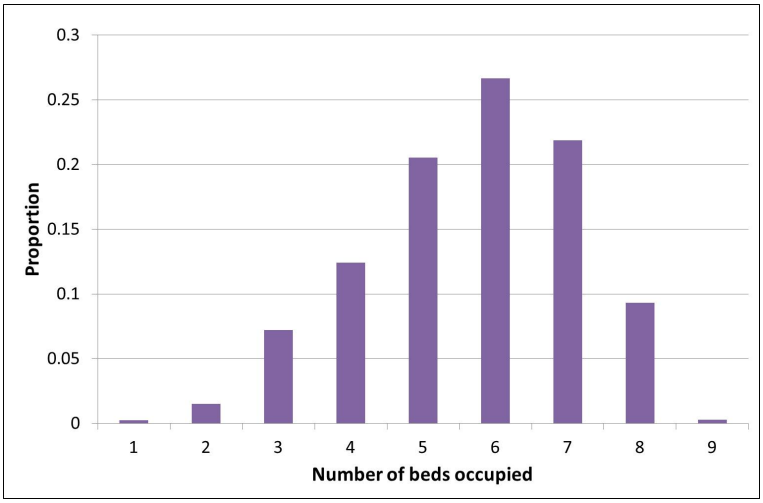


Figure 5.11: Bed occupancy frequency distribution

The frequency distribution from Figure 5.11 will be used later for comparison between the data and the mathematical model.

5.2.3.2 Royal Gwent

Royal Gwent is the larger of the two hospitals. Bed capacity was changed twice in the study period and therefore each period will be considered individually, since having 14 beds occupied during Period 1, when there were 14 beds available is not comparable to having 14 beds occupied in Period 3, when bed capacity was 16. The number of beds occupied at every hour in the study period is illustrated in Figure 5.12. The periods are separated by vertical lines.

Summary statistics of bed occupancy in each of the three periods in RG is presented in Table 5.7.

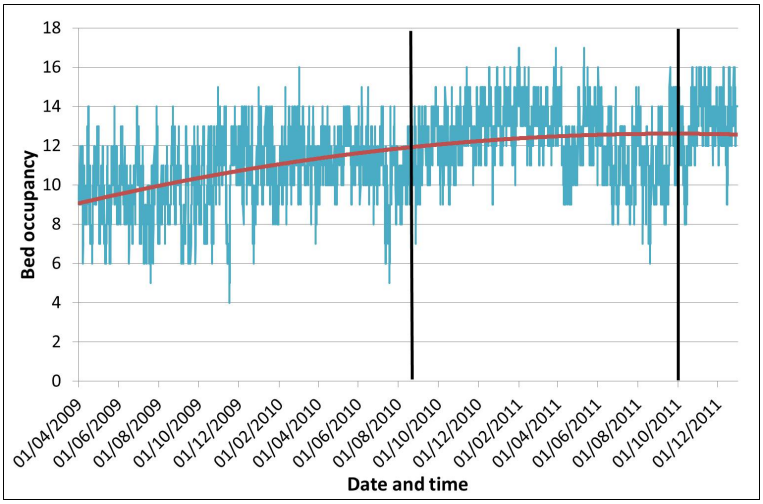


Figure 5.12: Hourly bed occupancy at RG (April 2009-December 2011)

Table 5.7: Summary statistics for bed occupancy in RG

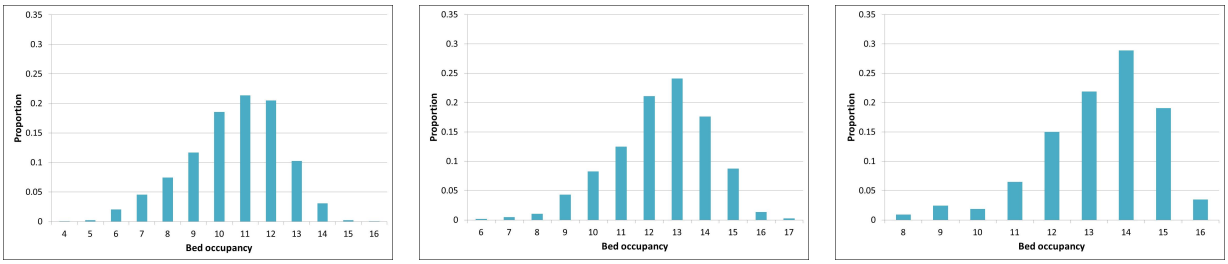
Summary statistic	**Period 1**	**Period 2**	**Period 3**	**Overall**
Mean	10.57	12.43	13.3	11.54
Median	11	13	14	12
Standard Deviation	1.84	1.74	1.57	2.06
Minimum	4	4	8	4
Maximum	16	17	16	17

The overall average bed occupancy is 11.54 with standard deviation of 2.06; however, the variability across the periods is significant. As the bed capacity is increased the mean bed occupancy is also increased. In Period 1, on average 10.57 beds were used on each day, giving a utilisation rate of 75.5%. In Period 2, the average is increased to 12.43 (utilisation 82.8%), and again in Period 3 to 13.3 (utilisation 83.1%). The red trend-line in Figure 5.12 also shows an increasing trend, especially on transition between Period 1 and Period 2. The number of beds occupied fluctuates up to 17 during busy periods and dropping to 4 at quiet times. An analysis reveals that on 0.26% of occasions in Period 1 and on 1.64% in Period 2, bed occupancy exceeded total bed capacity. It is observed that in Period 3 bed occupancy never exceeded the bed availability.

The data is investigated to check for any time dependencies. The data is split into the three periods for this purpose. Similar trends are observed in Period 1 and Period 2, the highest average bed occupancy on Wednesday (11.01 beds) in Period 1 and on Tuesday (12.86 beds) in Period 2 with the lowest average on Saturday (10.30 beds) in Period 1 and Friday (12.05 beds) in Period 2. Period 3 has a very different weekly trend: the average bed occupancy fluctuates between 14.01 on Saturday

and 12.44 on Wednesday. The hourly trends are examined for each of the period. The results are very similar across the whole study period; it appears that on average the day starts with the lowest average bed occupancy, it then steadily increases to reach its maximum around midday and then start to decrease again towards the evening.

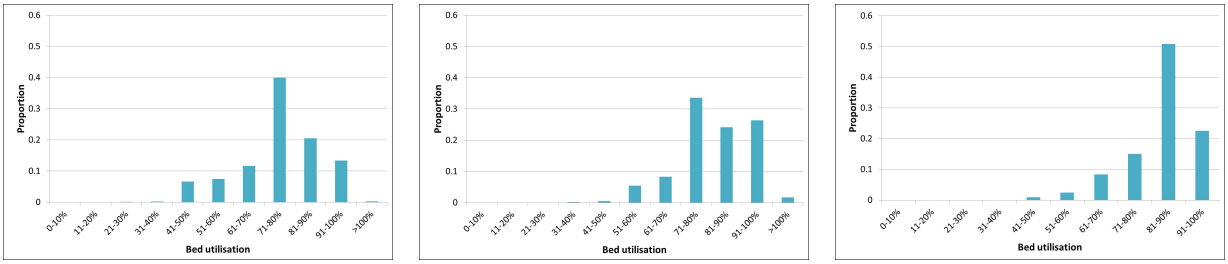
Consider the frequency distributions of the bed occupancies in the CCU during each period, presented in Figures 5.13a, 5.13b and 5.13c. The bed occupancy profiles look similarly across the three periods: small proportion of low bed occupancies, high probability of bed utilisation around 70-85% and low proportion of very high bed occupancies. The median value varies across the three periods, it starts with eleven, then increases to 13 and increases again to 14 in Period 3.



(a) Period 1 (b) Period 2 (c) Period 3

Figure 5.13: Bed occupancy frequency distribution for each period

This bed occupancy representation is not functional for later modelling purposes, therefore it is decided to examine utilisation rates or percentage distribution, i.e. on how many occasions are there $0-10\%, 11-20\%, 21-30\%$ etc. beds occupied. It is done in the following way: the possible number of beds occupied are divided by the bed capacity in each period and then grouped in eleven classes: $0-10\%, 11-20\%, \ldots, 90-100\%, > 100\%$. Consider the percentage frequencies displaying the bed occupancy proportion in the CCU during each period, presented in Figures 5.14a, 5.14b and 5.14c.



(a) Period 1 (b) Period 2 (c) Period 3

Figure 5.14: Bed occupancy frequency distribution for each period

The three frequency distributions are combined into one by taking the proportion of time each period lasted; Period 1 lasted 17 months; Period 2, 13 months and Period 3, 3 months. Each bar in

the overall bed occupancy percentage distribution is obtained in the following way:

$$P(x\% - y\%) = \sum_{i=1}^{3} P(\text{ Period } i) \times P(x\% - y\% \text{ in Period } i)$$

Figure 5.15 illustrates the overall bed utilisation frequencies. The overall utilisation rate is 79%, and a bed utilisation of $71 - 80\%$ has the highest probability of occurrence (35%). It is noteworthy that on 20% of occasions bed utilisation exceeds 90%, which means that the Unit is running at a very high occupancy level 20% of the time. The frequency distribution from Figure 5.15 will be later used for comparison between the data and the model.

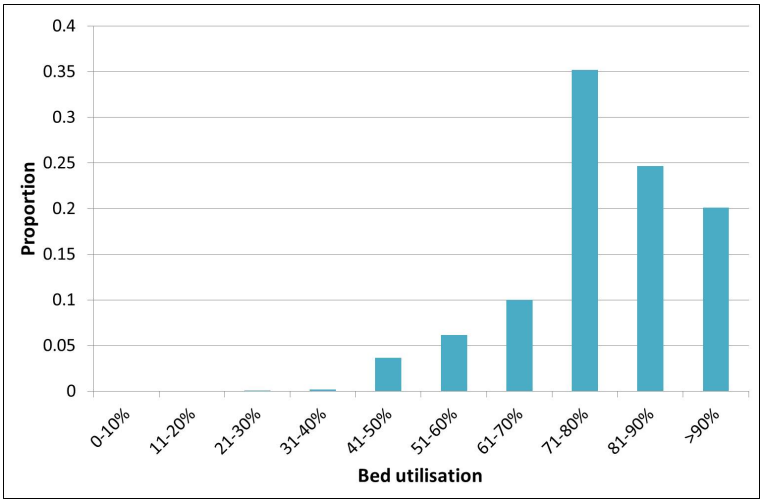


Figure 5.15: Overall bed occupancy frequency distribution at RG (April 2009-December 2011)

After inspection of the data sets it is noted that some patients were occupying beds even though they did not require life support any longer but, for example, because there was no free bed for them in an ordinary ward, and those patients were blocking CCU beds. Section 5.2.4 will investigate the occurrence of delayed discharges.

5.2.4 Delay to Discharge

Critical care services are very expensive, so it is very important to keep costs as low as possible without any compromise of the quality of care. This section examines the occurrence of delayed discharges.

Delay in discharge is defined as a difference between the moment the patient is ready for discharge and the actual discharge. Patients should be discharged from a CCU when the specialised care is no longer required. Sometimes, however, some issues cause a delay in a patient's discharge. The

factors contributing to the delay of discharge are unspecified in the data set; however, the main reason quoted by the hospital managers was unavailability of ward beds. Other quoted reasons included: medical complications, transport problems (transferring patient to another hospital) or lack of ward nursing staff with adequate skills. Some delays to discharge happen as a result of the 'human factor'; i.e. if there was no pressure on CCU beds, the CCU manager would make a decision to delay patient's discharge even though medically the patient was ready to be transferred to an ordinary ward. The term delayed discharge in the literature is also sometimes known by the term 'bed blocking'.

The reported estimates of the extent of delayed discharge vary between studies because of variation in methodology (e.g. restriction of study to particular ward or age groups). Coast, 1996 [29] estimated that 15-25% of all admissions and 60% of all bed days were potentially inappropriate. Victor *et al.*, 1993 [162] reported that 19% of those aged 65 and above in two inner hospitals were defined by staff as 'bed blockers', which represented 8% of all acute beds. However, the same study also indicated that delayed discharge was not a problem exclusively associated with older people, as they represented only about half of delayed patients. Healy *et al.*, 1999 [85] examined the extent of delayed discharge amongst patients in three hospital elderly care units in England, and analysed the factors associated with such delays using the conceptual model of individual and organisational factors (e.g. disagreements between health and social services).

Victor *et al.*, 2000 [161] in their study of three hospital elderly care units showed that among older patients in geriatric units, factors such as age and frailty are not causes for a delay in discharge. The authors concluded that the major cause of delay is organisational / administrative, which is compounded by constrained social services budgets.

Williams and Leslie, 2004 [167] examined the prevalence and reasons for delayed discharge in ICU. The majority (81%) of delays happened as a result of unavailable ward beds. Delays due to medical reasons only accounted for 8.5% of the patient delays from the ICU. The authors also showed that patients who were more severely ill on admission were more likely to be delayed and that most delays occurred during the weekend.

Maessen *et al.*, 2008 [116] designed a study to assess the influence of an ERAS (enhanced recovery after surgery) program on the proportion, appropriateness, and extent of delay in discharge. The degree of delay in discharge decreased significantly from a median of two days to a median of one day in the ERAS group.

Lim *et al.*, 2006 [112] observed that elderly patients are more likely to stay longer as inpatients compared with younger ones even after their acute medical problems have been resolved. They are

more prone to hospitalisation-related complications, like infections or worsening function.

The following two sections will examine the delay in discharge from the CCUs at NH and RG and what factors might influence patients' delays to discharge.

5.2.4.1 Nevill Hall

In the study period there were 1640 discharges from NH CCU, including 79% of patients who experienced any delay in their discharge. Patient discharge delay time ranged from 10 minutes to 21 days (mean, 9.25 hours; standard deviation, 20.25 hours; median, 3.83 hours). The delay times are grouped into 8-hourly time periods. The majority of delayed patients (71%) are discharged within eight hours of the decision to discharge and 84% within 24 hours.

Since there are 8 beds available, it gives a total of 8760 bed-days in the three year study period. The total delay to discharge time is 632 days, which implies that 7.21% of beds were blocked on each day. Given that there are 8 beds, implies that on average 0.58 beds are blocked per day, which seems relatively high for such a small CCU.

As said previously, critical care services are very expensive. The cost of an additional 632 bed-days in the study period cost the NHS approximately £1.1 million, calculated using the average daily cost of CCU bed as £1800 ([129]).

It is decided to investigate whether CCU bed occupancy had an influence on patient delays. Intuitively, if the Unit is busy, there is more pressure to free up beds, therefore the delay should be significantly lower than when bed utilisation is low. Using the data it is found on what day and what time each patient was ready to be discharged, simply by taking away the delay to discharge time from the actual discharge time. Then, it is checked how many beds are occupied in the time that patient was ready to be discharged and the mean delay was found for each bed occupancy. Figure 5.16 illustrates how the mean delay in discharge changes for different bed occupancies together with probabilities of a given occupancy (orange line with the scale on the right hand side).

As expected, the mean delay to discharge time is highest (on average 11 hours) when the Unit is not very busy. When bed utilisation reaches 75% the mean delay time starts to rapidly decrease. Delays decrease as occupancy increase to maximum capacity, which may reflect increasing discharge effort in order to ensure beds were free for new admissions.

The data is also investigated to check any time dependencies, i.e. to test whether the day of the week had an influence on the delay to discharge. Figure 5.17 demonstrates any existing time de-

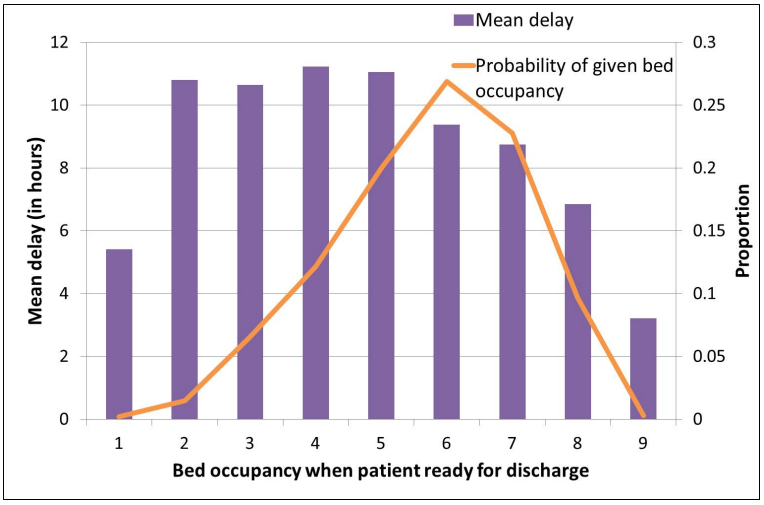


Figure 5.16: Mean discharge delay at NH for given bed occupancy when patient is ready for discharge

pendencies together with probabilities of the delay happening on each day of the week (scale on the right hand side).

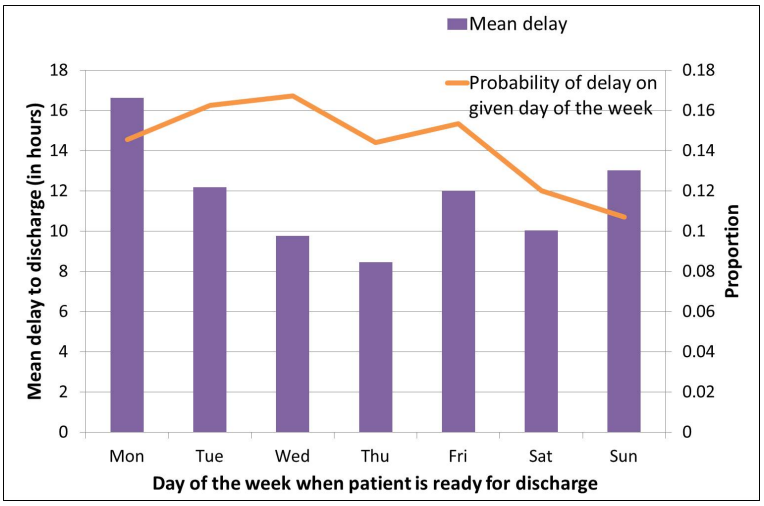


Figure 5.17: Mean discharge delay at NH for given day of the week when patient is ready for discharge

The most delays occurs on Wednesday (17% of all delays), closely followed up by Tuesday (16%) and Monday (15%), with longest delays on average (16.6 hours). Sunday has the lowest probability of delay to discharge (11%), but the average delay time was second highest (13 hours).

Other factors that could possibly be associated with postponement of discharge is patient age. As observed by Lim *et al.*, 2006 [112] elderly patients' discharges are more likely to be delayed. The relationship between age and delay time is investigated. All patients are grouped into 10-year age

groups; patients in the age group 71-80 accounts for most delays (23% of all delayed patients) and young patients group (younger than 20 years old) for only 6% of all delayed patients. Excluding the patients in age group below 20 and above 90, who in total accounted for only 6% of all admissions the longest average delay to discharge have patients in the age group 21-30 (15 hours) and the shortest patients in the age group 81-90 (8.3 hours). It can be concluded that elderly patients' discharges are more likely to be delayed; however, on average the delay time is below overall average.

5.2.4.2 Royal Gwent

In the analysed study period there were 2458 patients who were discharged from the CCU at RG, including 60% of patients who experienced any delay in their discharge, which compares favourably with NH, where 79% of patients were discharged with delay. Patients' discharge delay time ranged from 10 minutes to 16.2 days (mean, 20.3 hours; standard deviation, 35.6 hours; median, 4.5 hours). The delay times were grouped into 8-hourly time periods. The majority of delayed patients (35%) were discharged within 8 hours of the decision to discharge.

Recall, the number of available beds in RG has changed twice during the study period. For the first 20 months there were 14 beds, for the next 13 months 15 beds, and in the last three months 16 beds available; this gives the total of 15,929 bed-days in the three year study period. The total delay to discharge sums up to 2081 days. Each period is considered separately to check whether bed capacity has an influence on the discharge delay. Table 5.8 shows summary statistics and a few calculated variables.

Table 5.8: Summary statistic for discharge delay in each period in RG

Variable	**Period 1**	**Period 2**	**Period 3**
Mean	15.3 hours	25.3 hours	36 hours
Median	2.2 hours	7 hours	6.7 hours
Mode	0	0	0
Standard deviation	1.2 days	1.7 days	2.2 days
Minimum	0	0	0
Maximum	9.3 days	16.2 days	10.4 days
Percentage of delayed patients	54%	67%	75%
Total delay	761.5 days	900.4 days	301.7 days
Number of beds available	14	15	16
Number of bed-days	8522	5935	1472
Average percentage of blocked beds per day	8.94%	15.17%	20.5%
Mean number of blocked beds per day	1.25	2.28	3.28

Table 5.8 shows very interesting results: increasing bed capacity increases:

- the number of delayed patients (54% in Period 1, 67% in Period 2 and 75% in Period 3).
- the average delay time (15.3 hours in Period 1, 25.3 hours in Period 2 and 36 hours in Period 3).
- the percentage of blocked beds (8.94% in Period 1, 15.17% in Period 2 and 20.5% in Period 3) .

It is clear, that the bed capacity has an impact on discharge delays. It is examined whether the bed occupancy had an influence on patient delays. Figure 5.18 demonstrates how the mean delay in discharge changes for different bed occupancies for the whole period of three years, together with probabilities of given occupancy (red line with the scale on the right hand side).

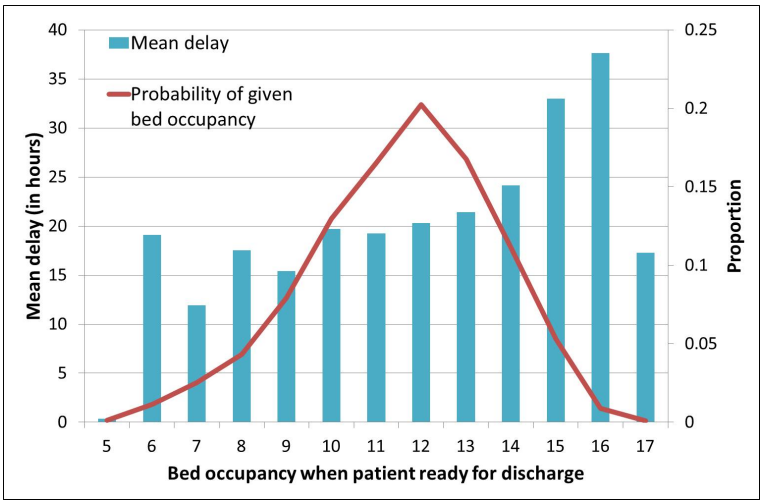


Figure 5.18: Mean discharge delay at RG for given bed occupancy when patient is ready for discharge

Surprisingly, Figure 5.18 shows very different results to the CCU at NH (Figure 5.17). As bed occupancy increases, the average discharge delay time increases. The highest mean delay is observed for bed occupancy of 16; however, the proportion of time that bed occupancy was 16 was very low (0.9%). Nevertheless, looking only at bed occupancies with probability of occurrence greater than 5% (i.e. bed occupancy between 9 and 15), an increasing trend is evident. Thus, the bed occupancy is not found to be a major factor in discharge decisions in RG.

The RG data is also investigated to check for any time dependencies. Figure 5.19 demonstrates any existing time dependencies together with probabilities of delay happening on each day of the week. Similarly as in NH, most delays occur on Wednesday (16% of all delays), closely followed up by Friday (15.6%). Sunday has the lowest percentage of delays to discharge (11%), but on average

longest delays (39.2 hours). The shortest delays on average were experienced on Thursday (29 hours).

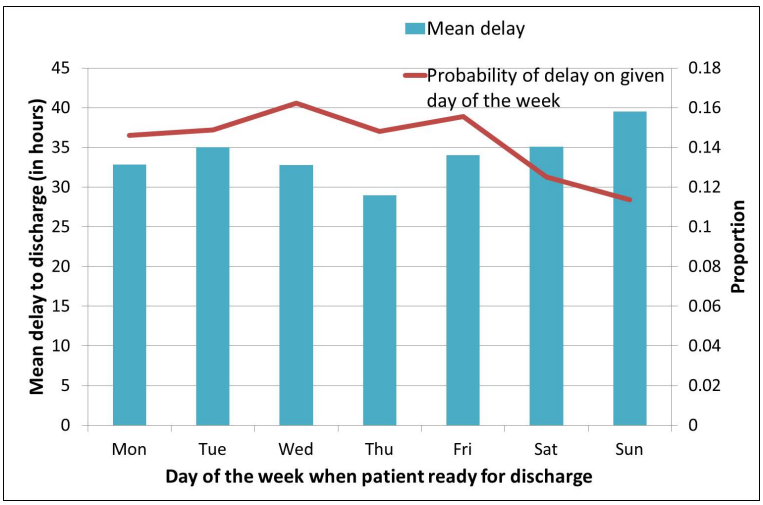


Figure 5.19: Mean discharge delay at RG for given day of the week when patient is ready for discharge

The last factor investigated is patient age, to check whether delay to discharge is related to it. As observed in NH, elderly patients' discharges are more likely to be delayed; it is found that patients in the age group 71-80 accounted for most delays (24% of all delayed patients) and young patients (younger than 20 years old) for only 2% of all delayed patients. Excluding the patients in age group below 20 and above 90, who in total account for only 3% of all admissions, the longest average delay to discharge is patients in the age group 41-50 (22.3 hours), closely followed up by patients in the age group 71-80 (22 hours), and the shortest patients in age group 21-30 (16.8 hours). It can be concluded that elderly patients' discharges are more likely to be delayed and delays are fairly significant.

5.2.4.3 Conclusions

Proactive management of early discharge planning with the attention on the changing care needs of the patient and better utilisation of ordinary ward beds is essential to reduce delays in discharge from CCUs. Reducing delays in the discharge process would free up beds for other admissions; which would benefit patients and would result in cost savings for the hospitals. Delays in discharge have cost implications; additional nursing time is needed. By reducing delayed discharge a more cost-effective health care system can be provided.

Section 5.2.4 determined the occurrence of discharge delays in the CCUs discharge and identified some factors that influence the postponement. It was shown that bed occupancy in NH could

sway discharge decisions, however this was not observed in RG. Also, there were visible time dependencies in both hospitals, i.e. day of the week has an impact on discharge delay. The study also confirmed the work of Lim *et al.*, 2006 [112] which showed that elderly patients' discharges are more likely to be delayed.

5.3 Conclusions

The main aim of this chapter has been to investigate patient flow including the admission process and length of stay profile in the CCUs at the Nevill Hall and the Royal Gwent Hospitals. Adequate distributions which accurately described admission processes and duration of stay in the Units were provided. An interesting conclusion was that increasing bed capacity did not increase the number of admissions, but increased patient duration of stay, which then caused greater bed utilisation rates. Other important factor influenced by increased bed capacity was delay to discharge.

Information provided in this chapter will be used in Chapter 6 to consider aspects of theoretical applications of mathematical modelling of the CCUs.

Chapter 6

Mathematical Modelling of the Nevill Hall and the Royal Gwent Critical Care Units

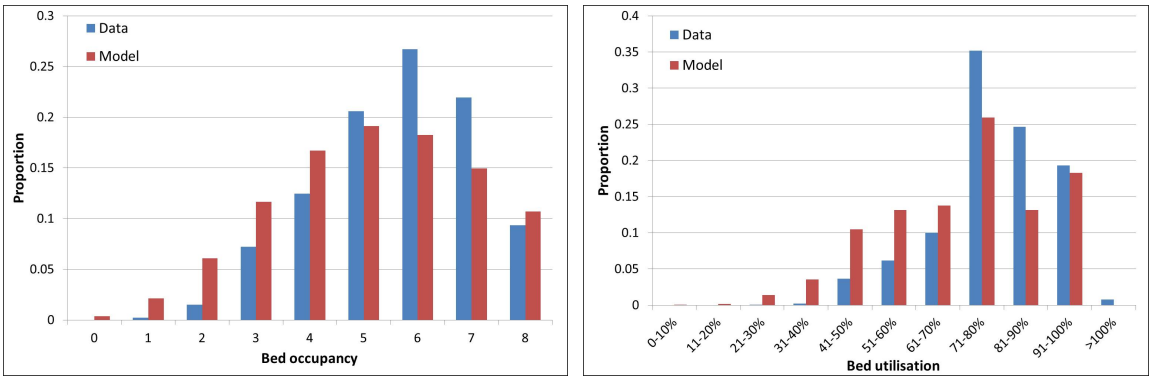
6.1 Introduction

The purpose of this chapter is to develop and validate a mathematical model of bed occupancies at Critical Care Units (CCU) at Royal Gwent and Nevill Hall. In Section 6.4, sensitivity of the model will be tested by a few 'what if' scenarios, including transfer of patients and beds from one CCU to other. In Section 6.5, sensitivity of the model will be explored through changes to the combined Unit size.

6.2 The Queueing Model

Any queueing model is dependent on the accurate assessment of three variables: arrival rate, service time and the number of channels in the system. Bed occupancy is initially modelled using Erlang's Loss Formula for the $M/M/c/c$ queue, which previously proved to accurately model the CCU at the University Hospital of Wales in Cardiff (Section 3.3). The analytical results for the Nevill Hall (NH) and the Royal Gwent (RG) however, do not provide a satisfactory fit, as can be seen in Figures 6.1a and 6.1b. For NH the model shows a lower frequencies of bed occupancy values five, six and seven, and higher frequencies at both ends of the distribution. For RG the model underestimates bed utilisation above 70%, and overestimates low bed occupancy. The possible explanation for these discrepancies is that the system does not work as "smoothly" as its mathematical model; therefore, a more complex bed occupancy model has to be considered. It is suggested, that the number of admissions might be dependent on current bed occupancy levels; therefore, in this study queueing theory is used to develop a new mathematical model of patient flow with cut-off points. The model allows for different admission rates for various bed occupancy levels, hence is an extension to the

model described in Sections 3.3.6 and 3.3.7. The cut-off points are not strict guidelines to hospital managers, but they reflect on behavioural aspects.



(a) NH (b) RG

Figure 6.1: Initial bed occupancy model for each CCU

A system, in which the mean service rate depends on the state of the system was described in Stewart, 2009 [152]. The case in which the server works at rate μ_1 until there are k customers in the system, at which point it changes to a different rate μ_2 was considered. The hospital managers might not have great control of patients LoS, but they have control over patient admissions. Therefore this part of the project concentrates on state-dependent arrivals.

Section 1.4 considered queueing systems with variable arrival rates. Gong *et al.*, 1992 [60] considered the $M/G/1$ queue with queue-length dependent arrival rates. Courtois and Georges, 1971 [37] generalised the $M/G/1$ queueing process by considering the arrival and the service rates as being arbitrary functions of the current number of customers in the system. The existing literature on state-dependent queueing systems mainly considers systems with a single server, while a CCU is a multi-channel system.

The CCUs are modelled as a multi-channel, single-stage system with identical parallel servers where queueing is not allowed (loss queue). If all beds are occupied, then an arriving patient is transferred to a different hospital or to an alternative ward within the hospital. Those patients are referred to as 'turn away' or rejected. Each CCU bed is treated as one server and a $FIFO$ (first-in-first-out) queueing discipline is assumed. Let P_n be the steady-state probability that there are n patients in the system and let:

k_1, k_2	cut-off number 1, 2 $(k_2 > k_1)$
λ_a	the arrival rate if bed occupancy $\leq k_1$
λ_b	the arrival rate if $k_1 <$ bed occupancy $\leq k_2$
λ_c	the arrival rate if bed occupancy $> k_2$
μ	the service rate
c	the number of channels (beds)

A diagrammatic representation is given in Figure 6.2.

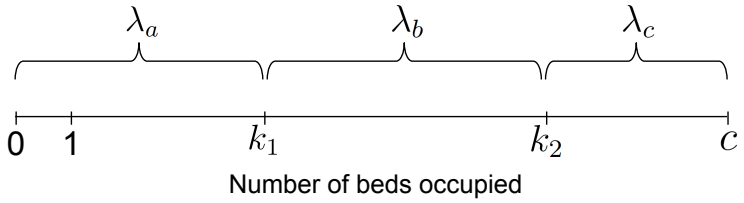


Figure 6.2: State dependent queueing model

The system with c service channels available can be described by the set of the differential-difference equations:

(1) For $n = 0$:

$$\begin{aligned} P_0(t+\delta t) =& P_0(t)(1-\lambda_a \delta t) \\ &+ P_1(t)(1-\lambda_a \delta t)\mu \delta t + o(\delta t) \end{aligned}$$

(2) For $1 \leq n \leq (k_1 - 1)$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)(1-\lambda_a \delta t)(1-n\mu \delta t) \\ &+ P_{n-1}(t)\lambda_a \delta t\,(1-(n-1)\mu \delta t) \\ &+ P_{n+1}(t)(1-\lambda_a \delta t)(n+1)\mu \delta t + o(\delta t) \end{aligned}$$

(3) For $n = k_1$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)(1-\lambda_b \delta t)(1-n\mu \delta t) \\ &+ P_{n-1}(t)\lambda_a \delta t\,(1-(n-1)\mu \delta t) \\ &+ P_{n+1}(t)(1-\lambda_b \delta t)(n+1)\mu \delta t + o(\delta t) \end{aligned}$$

(4) For $(k_1+1) \leq n \leq (k_2-1)$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)(1-\lambda_b\delta t)(1-n\mu\delta t) \\ &+ P_{n-1}(t)\lambda_b\delta t\,(1-(n-1)\mu\delta t) \\ &+ P_{n+1}(t)(1-\lambda_b\delta t)(n+1)\mu\delta t + o(\delta t) \end{aligned}$$

(5) For $n = k_2$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)(1-\lambda_c\delta t)(1-n\mu\delta t) \\ &+ P_{n-1}(t)\lambda_b\delta t\,(1-(n-1)\mu\delta t) \\ &+ P_{n+1}(t)(1-\lambda_c\delta t)(n+1)\mu\delta t + o(\delta t) \end{aligned}$$

(6) For $(k_2+1) \leq n \leq (c-1)$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)(1-\lambda_c\delta t)(1-n\mu\delta t) \\ &+ P_{n-1}(t)\lambda_c\delta t\,(1-(n-1)\mu\delta t) \\ &+ P_{n+1}(t)(n+1)\mu\delta t + o(\delta t) \end{aligned}$$

(7) For $n = c$:

$$\begin{aligned} P_n(t+\delta t) =& P_n(t)(1-n\mu\delta t) \\ &+ P_{n-1}(t)\lambda_c\delta t\,(1-(n-1)\mu\delta t) + o(\delta t) \end{aligned} \tag{6.1}$$

The steady-state equations may be written in the form shown below:

(1) For $n = 0$:

$$\mu P_1 = \lambda_a P_0$$

(2) For $1 \leq n \leq (k_1-1)$:

$$(\lambda_a + n\mu)P_n = \lambda_a P_{n-1} + (n+1)\mu P_{n+1}$$

(3) For $n = k_1$:

$$(\lambda_b + n\mu)P_n = \lambda_a P_{n-1} + (n+1)\mu P_{n+1}$$

(4) For $(k_1+1) \leq n \leq (k_2-1)$:

$$(\lambda_b + n\mu)P_n = \lambda_b P_{n-1} + (n+1)\mu P_{n+1}$$

(5) For $n = k_2$:

$$(\lambda_c + n\mu)P_n = \lambda_b P_{n-1} + (n+1)\mu P_{n+1}$$

(6) For $(k_2 + 1) \leq n \leq (c - 1)$:

$$(\lambda_c + n\mu)P_n = \lambda_c P_{n-1} + (n+1)\mu P_{n+1}$$

(7) For $n = c$:

$$n\mu P_n = \lambda_c P_{n-1}$$

Theorem 6.2.1. *The bed occupancy probabilities with two cut-off points k_1 and k_2, where $k_2 > k_1$ which are dependent on bed occupancy levels (as described by Figure 6.2) are given by:*

$$P_n = \begin{cases} \frac{1}{n!}{\theta_a}^n P_0 & \text{if } 0 \leq n \leq k_1 \\ \frac{1}{n!}{\theta_a}^{k_1}{\theta_b}^{n-k_1} P_0 & \text{if } (k_1+1) \leq n \leq k_2 \\ \frac{1}{n!}{\theta_a}^{k_1}{\theta_b}^{k_2-k_1}{\theta_c}^{n-k_2} P_0 & \text{if } (k_2+1) \leq n \leq c \end{cases}$$

where

$$\theta_a = \frac{\lambda_a}{\mu} \qquad \text{and} \qquad \theta_b = \frac{\lambda_b}{\mu} \qquad \text{and} \qquad \theta_c = \frac{\lambda_c}{\mu}$$

and

$$P_0 = \frac{1}{1 + \sum_{r=1}^{k_1} \frac{1}{r!}\left(\frac{\lambda_a}{\mu}\right)^r + \left(\frac{\lambda_a}{\mu}\right)^{k_1} \sum_{r=k_1+1}^{k_2} \frac{1}{r!}\left(\frac{\lambda_b}{\mu}\right)^{r-k_2} + \left(\frac{\lambda_a}{\mu}\right)^{k_1}\left(\frac{\lambda_b}{\mu}\right)^{k_2-k_1} \sum_{r=k_2+1}^{c} \frac{1}{r!}\left(\frac{\lambda_c}{\mu}\right)^{r-k_2}}$$

Proof.
The proof is similar to that of Theorem 3.3.1, it is done using an inductive argument, and is in Appendix D. □

The service rate parameter μ is taken from the data. The next task is to determine values for the parameters $k_1, k_2, \lambda_a, \lambda_b, \lambda_c$ which would give the closest representation of real-life bed occupancy probabilities. There are no written rules for managers to determine patient admission automatically; these decisions are subject to the judgement of the critical care consultant. The approach chosen to model this decision-making process requires defining a set of rules to determine arrival rates, which depend on bed occupancy levels. Since there are no official rules as to when to decrease arrival rates, it is impossible to obtain this information from the data. It is addressed by defining an optimization problem aimed at matching the mathematical model of bed occupancy with the data. This is achieved using the optimisation tool: *Evolver*.

Evolver is a software package build in Excel that allows users to tackle a wide variety of optimisation problems using a genetic algorithm ([54]). Applications written with *Evolver* can contain large numbers of adjustable cells and it can handle complex, non-linear problems, while *Solver* fails for large non-linear problems. In a genetic algorithm, each possible solution to a given problem becomes an independent 'organism' that can 'breed' with other organisms. The process is described below ([49]):

1. The *Evolver* engine randomly generates many organisms (possible solutions), and calculates the result each organism it produces. This entire 'population' of organisms is ranked from best to worst.

2. The genetic algorithm engine then selects good organisms and swaps their variables (genes) using crossover and mutation to produce 'offspring'. If offspring do not produce a good result, two more parents are selected.

3. If the offspring organism is good, it is re-inserted into the population.

As *Evolver* repeats steps 2 and 3, the population 'evolves' increasingly producing better solutions.

The mathematical program used to find model parameters is described below:

Minimise:

$$\sum_{n=0}^{c} (\text{data}P_n - \text{model}P_n)^2$$

Subject to:

$$0 < k_1 < k_2 < c$$

$$k_1, k_2 \in \mathbb{N}$$

$$\lambda_a, \lambda_b, \lambda_c \in \mathbb{R}_+$$

$$0.9\lambda \leq \Lambda \leq 1.1\lambda$$

where:

$$\Lambda = \sum_{n=1}^{k_1} \lambda_a P_n + \sum_{n=k_1+1}^{k_2} \lambda_b P_n + \sum_{n=k_2+1}^{c} \lambda_c P_n$$

Note, if $k_1 = k_2$, $k_1 = 0$ or $k_2 = c$ then the problem reduces to one cut-off model, which was considered, and did not provide a good fit. The final constraint is added to make the model as accurate as possible, it is decided to allow the overall model arrival rate (Λ) to be within 10% of the actual admission rate.

6.3 Results of the Mathematical Model

Having developed a mathematical model of the CCUs based on actual data and parameters obtained from the solution to the mathematical program, it is necessary to investigate whether the new model provides a good bed occupancy fit. This section displays a brief outline of the results for each of the hospitals.

Evolver is run to find parameter values that would provide the best bed occupancy fit at NH. The objective function, sum of squared differences of the model and data bed occupancies is equal to 0.004, and is minimised for the following set of parameters:

Table 6.1: Parameter values

Parameter	**Value**
λ_a	0.644
λ_b	1.916
λ_c	1.263
k_1	2
k_2	6

The found parameters indicate that the hospital admits on average 0.644 patients per day when bed occupancy is 2 or lower. It might seem counter-intuitive, however the probability of having 2 or less beds occupied was very low, suggesting there were no doctors or nurses, hence low rate of arrivals. When there are between 3 and 6 beds occupied the arrival rate is increased to 1.916 and then, if bed occupancy exceeds 6 the arrival rate decreases to 1.263. The overall arrival rate is 1.647, which is exactly 10% higher than the observed admission rate (1.498). The other factor that suggests whether the model provides a good fit is the utilisation rate. The utilisation rate, which is calculated in the following way: $\frac{\sum_{n=1}^{c} nP_n}{c} = 0.67$ compares favourably with the observed rate of 0.71.

The analytical results are now compared with the data in Figure 6.3, which shows close agreement in bed occupancy levels when comparing the model results with the data. Recall, the total bed capacity is 8. As shown in Section 5.2.3.1 on some occasions the number of beds occupied exceeded the total capacity; in the mathematical model the number of server channels was chosen to be 8, therefore the data was standardised so that the maximum number of beds is 8.

The model was also run without the overall admission rate constraint and with decreased acceptance range for the overall admission rate from 10% to 5% and then to 1%. Table 6.2 shows parameter values for the three cases.

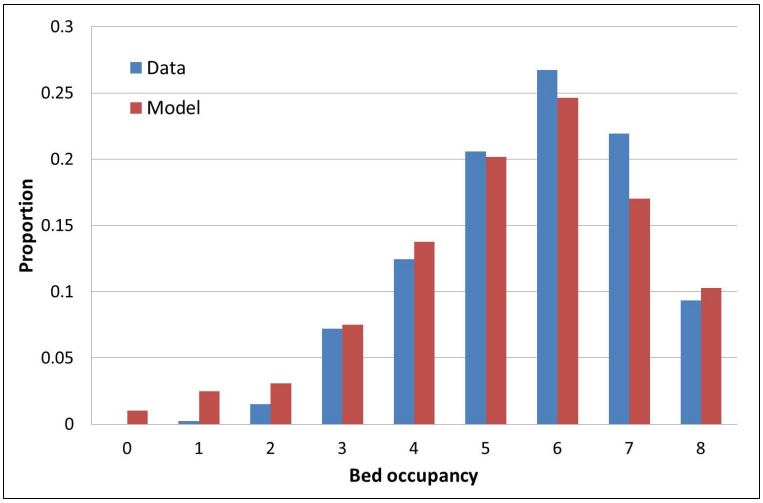


Figure 6.3: Comparison of analytical results with the NH data

Table 6.2: Parameter and variable values

Parameter/ Variable name	**No overall admission rate constraint**	**Overall admission rate within 5%**	**Overall admission rate within 1%**
λ_a	2.08	1.104	0.493
λ_b	1.493	1.865	1.781
λ_c	0.883	1.251	1.253
k_1	6	3	2
k_2	7	6	6
sum of squares	0.0002	0.007	0.009
overall admission rate	1.84 (23% incr.)	1.57 (5% incr.)	1.51 (1% incr.)

The model without the overall admission rate constraint provides a better bed occupancy fit; however, a 23% increase in the total admission rate is not accepted. Decreasing the acceptance range for the overall admission rate decreases the goodness of fit.

The model was also run with higher number of cut-off points (three and four); however, the results are not much better; therefore, for the simplicity and to avoid over-fitting of the model it is decided to accept the model with two cut-off points.

Attention is now given to the RG hospital, which requires slightly more care since the model has to be fitted for each of the three periods, and then combined into one with adequate probabilities. *Evolver* is run to find parameter values that would provide the best bed occupancy fit for each period. The objective function, sum of squared differences between the model and the data bed

occupancy probabilities, is 0.000311 in Period 1, 0.000033 in Period 2 and 0.001113 in Period 3 and is minimised for the set of parameters in Table 6.3.

Table 6.3: Parameter values

Parameter	**Value in Period 1**	**Value in Period 2**	**Value in Period 3**
λ_a	3.081	3.327	3.761
λ_b	2.462	2.628	1.754
λ_c	1.377	1.582	0.506
k_1	11	12	14
k_2	12	13	15

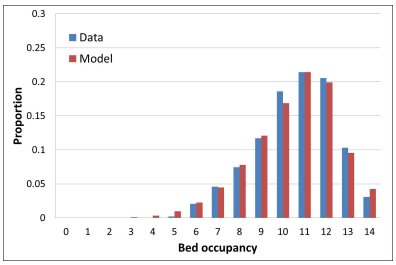
The obtained parameters indicate, as expected, that less patients on average are admitted when there are more beds occupied in the CCU, i.e. arrival rates are decreased as bed occupancy is increased. The overall arrival rates obtained from the model are compared with the observed rates and are illustrated in Table 6.4

Table 6.4: Comparison of overall arrival rates

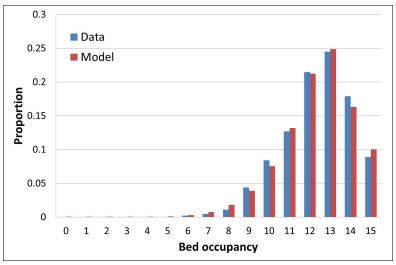
Period	**Data**	**Model**	**Increase**
Period 1	2.3026	2.5329	10%
Period 2	2.1722	2.277	4.87%
Period 3	2.1739	2.3913	10%

The other factors that indicate whether the model provides a reasonably good fit is the utilisation rate. The utilisation rate obtained from the model in Period 1 is 74.93%, which compares favourably with the data rate of 75.41%; similarly in Period 2, 82.26% with 82.43%; and in Period 3, 82.55% with 83.10%

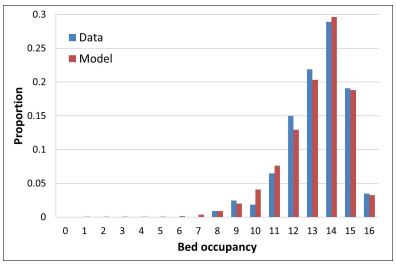
Figures 6.4a, 6.4b and 6.4c show how analytical results compare with the actual data in each period.



(a) Period 1



(b) Period 2



(c) Period 3

Figure 6.4: RG bed occupancy model fit for each period

In order to get the overall bed occupancy profile, bed occupancy proportions in each period are grouped into ten classes: $0-10\%, 11-20\%, \ldots, 91-100\%$, and the proportion that each period lasted for is incorporated in the following way:

$$\begin{aligned} P_{\text{overall}}(10s\% < n \leq 10(s+1)\%) = {} & P(\text{Period 1}) \times P_{\text{Period 1}}(10s\% < n \leq 10(s+1)\%) \\ & + P(\text{Period 2}) \times P_{\text{Period 2}}(10s\% < n \leq 10(s+1)\%) \\ & + P(\text{Period 3}) \times P_{\text{Period 3}}(10s\% < n \leq 10(s+1)\%) \end{aligned}$$

where $s \in [0, \ldots, 9]$. Figure 6.5 illustrates the overall bed occupancy profile with the model fit, which provides very accurate representation of the actual RG data.

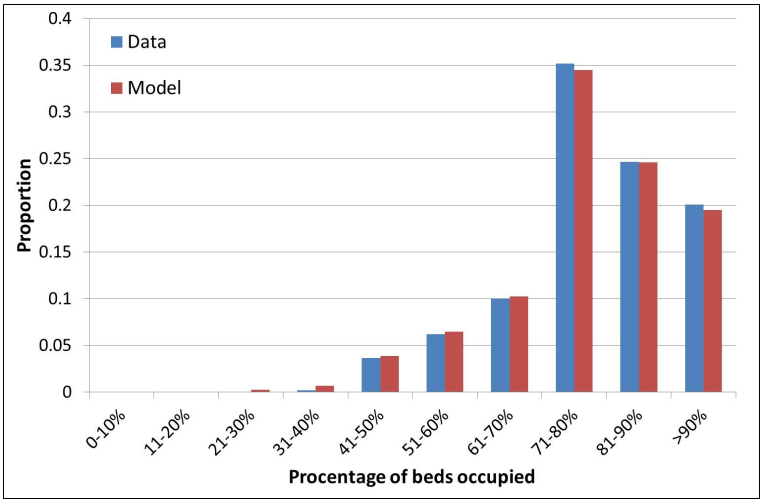


Figure 6.5: Comparison of analytical results with the RG data

Having obtained the model that accurately describes activities in both CCUs, it is now possible to test a few 'what if' scenarios, which will be described in Section 6.4. The model obtained herein will be referred to as the base model in Section 6.4.

6.4 'What if' Scenarios

The previous section detailed the validation of the base model; however, no alternations were made: arrival rates remained unchanged. This section will examine the effect of implementing some new policies regarding transfers of patients and beds between hospitals. It has been mentioned previously that sometimes the Unit is busy and patients need to be admitted to an ordinary ward, or if patients are very ill and need specialised care, they will be transferred to another CCU.

6.4.1 Transfer of Patients Between Hospitals

This section will investigate the consequence of transferring a percentage of patients from one CCU to the other. One of the hospitals might specialise in, for example, care of patients who undergo a transplant operation and all patients from this Local Health Board will be transferred to that CCU. However, the hospital that is receiving extra patients should not be experiencing overcrowding on too many occasions. The measures that will be taken under investigation are: the probability of a patient being rejected in both hospitals and the overall throughput of patients. The scenario, where throughput is higher and probability of rejection is lower from the base model will be recommended to the hospital managers.

If patients are transferred, for example form NH to RG, they will be under the care of RG and hence it is assumed that those patients' length of stay characteristic will be under RG, i.e. the hospital that the patient is receiving care in.

Recall, RG bed occupancy model is a combination of three models, due to the bed capacity change. As a result, the overall combined RG arrival rate is calculated as follows:

$$\begin{aligned}
\Lambda_{RG} &= P(\text{Period 1}) \times \lambda_{\text{Period 1}} + P(\text{Period 2}) \times \lambda_{\text{Period 2}} + P(\text{Period 3}) \times \lambda_{\text{Period 3}} \\
&= P(\text{Period 1}) \times \left(\sum_{n=1}^{k_{1\,\text{Period1}}} \lambda_{a_{\text{Period1}}} P_n + \sum_{n=k_{1\,\text{Period1}}+1}^{k_{2\,\text{Period1}}} \lambda_{b_{\text{Period1}}} P_n + \sum_{n=k_{2\,\text{Period1}}+1}^{c_{\text{Period1}}} \lambda_{c_{\text{Period1}}} P_n \right) \\
&+ P(\text{Period 2}) \times \left(\sum_{n=1}^{k_{1\,\text{Period2}}} \lambda_{a_{\text{Period2}}} P_n + \sum_{n=k_{1\,\text{Period2}}+1}^{k_{2\,\text{Period2}}} \lambda_{b_{\text{Period2}}} P_n + \sum_{n=k_{2\,\text{Period2}}+1}^{c_{\text{Period2}}} \lambda_{c_{\text{Period2}}} P_n \right) \\
&+ P(\text{Period 3}) \times \left(\sum_{n=1}^{k_{1\,\text{Period3}}} \lambda_{a_{\text{Period3}}} P_n + \sum_{n=k_{1\,\text{Period3}}+1}^{k_{2\,\text{Period3}}} \lambda_{b_{\text{Period3}}} P_n + \sum_{n=k_{2\,\text{Period1}}+1}^{c_{\text{Period3}}} \lambda_{c_{\text{Period3}}} P_n \right)
\end{aligned}$$

The overall NH arrival rate is as follows:

$$\Lambda_{NH} = \sum_{n=1}^{k_{1_{NH}}} \lambda_{a_{NH}} P_n + \sum_{n=k_{1_{NH}}+1}^{k_{2_{NH}}} \lambda_{b_{NH}} P_n + \sum_{n=k_{2_{NH}}+1}^{C_{NH}} \lambda_{c_{NH}} P_n$$

As an illustration, proportions of 10%, 20% and 50% of patients will be transferred, firstly from RG to NH. The new increased NH parameters are:

$$\lambda_{a_{NH}}^{new} = \lambda_{a_{NH}} + x\%\Lambda_{RG}, \lambda_{b_{NH}}^{new} = \lambda_{b_{NH}} + x\%\Lambda_{RG} \text{ and } \lambda_{c_{NH}}^{new} = \lambda_{c_{NH}} + x\%\Lambda_{RG}.$$

The new decreased RG parameters for each of the periods are:

$$\lambda_{a_{RG}}^{new} = (1 - x\%)\lambda_{a_{RG}}, \lambda_{b_{RG}}^{new} = (1 - x\%)\lambda_{b_{RG}} \text{ and } \lambda_{c_{RG}}^{new} = (1 - x\%)\lambda_{c_{RG}}.$$

The parameter values of k_1 and k_2 stay unchanged. The bed occupancy probabilities are calculated using Theorem 6.2.1. The examined measures are: the throughput and the probability of rejection, $P_r = P_c$, for each of the hospitals separately. Then the throughput at RG is added to the throughput at NH; similarly the sum of the rejection probabilities are taken.

Similarly, the proportion of NH patients are now transferred to RG. The new increased arrival rate parameters for RG are:

$$\lambda_{a_{RG}}^{new} = \lambda_{a_{RG}} + x\%\Lambda_{NH}, \lambda_{b_{RG}}^{new} = \lambda_{b_{RG}} + x\%\Lambda_{NH} \text{ and } \lambda_{c_{RG}}^{new} = \lambda_{c_{RG}} + x\%\Lambda_{NH}.$$

Correspondingly, the decreased NH parameters are:

$$\lambda_{a_{NH}}^{new} = (1 - x\%)\lambda_{a_{NH}}, \lambda_{b_{NH}}^{new} = (1 - x\%)\lambda_{b_{NH}} \text{ and } \lambda_{c_{NH}}^{new} = (1 - x\%)\lambda_{c_{NH}}.$$

The parameter values of k_1 and k_2 stay unchanged.

Figure 6.6 illustrates how throughput in each hospital changes for each scenario. Each vertical bar represents throughput at RG (blue) and NH (purple) with the total for each scenario given at the top of each bar. Scenario labelled, for example, 'NH + 10% from RG' means that 10% of RG admissions are transferred to NH. The first three scenarios show how throughput changes when percentage of patients are transferred from RG to NH. The following three demonstrate changes to the system when a percentage of NH admissions are transferred to RG. Unsurprisingly, the throughput at hospital with reduced admissions decreases as the percentage of patients transferred increases.

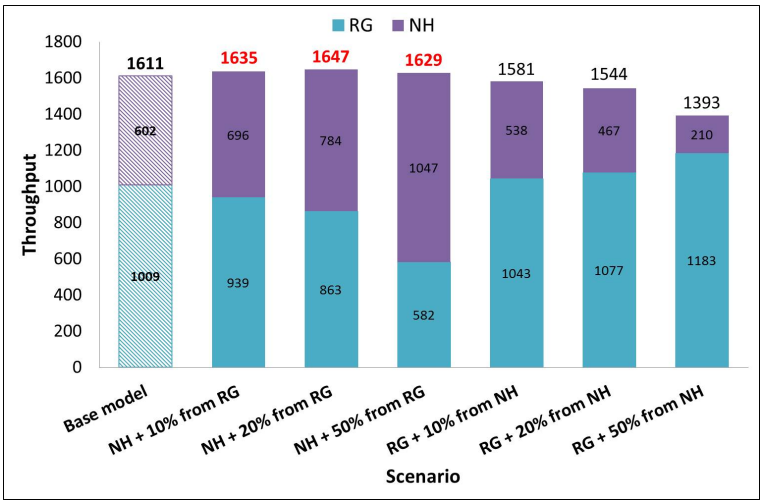


Figure 6.6: Throughput for each of the six scenarios

Consequently, the throughput at the hospital that is admitting extra patients increases as the percentage of transfers increases. The total throughput for the base model is 1611 patients per year; and the total is marked in red for scenarios with throughput higher than 1611. As it can be seen, the throughput is greater when patients are transferred from RG to NH (the first three scenarios), with the corresponding throughput of 1635 for 10%, 1647 for 20% and 1629 for 50% of extra transferred patients.

Increasing the throughput might increase the probability of a patient being rejected or equivalently, the probability of all beds being full. It is therefore necessary to investigate how this measure changes for each of the given scenarios. Figure 6.7 demonstrates how the probability of rejection changes.

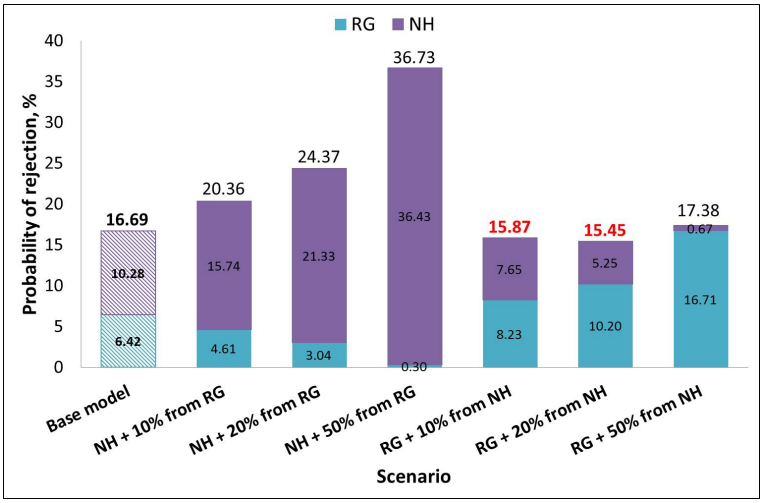


Figure 6.7: Probability of rejection for each of the six scenarios

The total probability of all beds being full for the base model is equal to 16.69%. Lower probabilities are obtained for the scenario where 10% or 20% of patients are transferred from NH to RG. Since RG is the bigger of the two hospitals it can deal better with the influx of patients. Also, if the bigger hospital admits extra patients, the probability of rejection increases steadily (8.23%, 10.20%, 16.71%) opposed to a rapid increase in rejection if the smaller hospital receiving extra admissions (15.74%, 21.33%, 36.43%).

Based on the two decision factors: the total throughput and the probability of rejection, improvements in the whole system can be made if given scenario increased the throughput and decreased the probability of rejection. As seen, total throughput was greater if a proportion of patients were transferred from RG to NH; however, if the decision is based on the lowest rejection proportion it is better to allow extra 10% or 20% of patients from NH to RG. To conclude, there does not exist a scenario that only allowed patients' transfers that would improve the system.

To summarise, the total throughput is higher if a percentage of patients are transferred from the bigger to the smaller CCU; however, this implies a significant increase of the probability of rejection.

The next 'what if' scenario will investigate the effect of patients' transfers, but this time also beds are allowed to be transferred.

6.4.2 Transfer of Patients and Beds Between Hospitals

This section will investigate the impact of transferring a percentage of patients from one CCU to the other, along with some beds. It is assumed that each of the hospitals have physical space for extra beds. Again, it is assumed that patients obtain length of stay characteristic of a CCU that they are receiving treatment in.

The measures that will be considered are again: the probability of a patient being rejected, due to insufficient number of beds, in both hospitals and the overall throughput of patients. The scenario, where throughput is higher and probability of rejection is lower from the base model will be recommended to the hospital managers.

If beds are transferred the following procedure is applied: assume one bed is transferred from NH to RG, then the number of available beds is decreased in NH to 7 and the bed capacity in RG is increased by 1. Since the overall RG model is a combination of three models, the number of beds is incremented in each period, resulting in 15 beds in Period 1, 16 beds in Period 2 and 17 beds in Period 3. The cut off points are now: $\widehat{k_1} = \left\lfloor \frac{\text{original } k_1}{\text{original bed capacity}} \times \text{new bed capacity} \right\rceil$ and similarly $\widehat{k_2} = \left\lfloor \frac{\text{original } k_2}{\text{original bed capacity}} \times \text{new bed capacity} \right\rceil$. The patients' transfer procedure is the same as described in the previous section.

As an illustration, it is assumed that up to 4 beds are allowed to be transferred; however, the model can easily be extended to a greater number. The proportion of patients that will be transferred is: 0%, 10%, 20% and 50%; again, this can be easily altered. In total, 32 scenarios are performed and Table 6.5 provides results of the throughput.

The scenario labelled, for example, 'RG +2 beds (20% from NH)' means that RG receives two beds from NH and an additional 20% of admissions are transferred from NH. The overall throughput in 17 scenarios was higher than the base model; those scenarios are highlighted in red in Table 6.5. The highest throughput of 1662 is observed for the scenario where three beds are transferred from NH to RG, but no patient transfer take place. The lowest (1415) is experienced for the scenario of transferring one bed and 50% of patients from NH to RG.

Table 6.5: Throughput results for each of the 32 scenarios

Scenario	**Throughput**		
	RG	**NH**	**Overall**
Base model	1009	602	**1611**
NH +1 bed (0% from RG)	979	626	1605
NH +1 bed (10% from RG)	914	721	1635
NH +1 bed (20% from RG)	842	807	1649
NH +1 bed (50% from RG)	577	1066	1643
RG +1 bed (0% from NH)	1036	569	1605
RG +1 bed (10% from NH)	1071	508	1579
RG +1 bed (20% from NH)	1107	441	1548
RG +1 bed (50% from NH)	1214	201	1415
NH +2 beds (0% from RG)	946	584	1530
NH +2 beds (10% from RG)	885	714	1599
NH +2 beds (20% from RG)	818	815	1633
NH +2 beds (50% from RG)	569	1082	1651
RG +2 beds (0% from NH)	1060	588	1648
RG +2 beds (10% from NH)	1098	520	1618
RG +2 beds (20% from NH)	1134	447	1581
RG +2 beds (50% from NH)	1244	201	1445
NH +3 beds (0% from RG)	955	582	1537
NH +3 beds (10% from RG)	891	708	1599
NH +3 beds (20% from RG)	821	804	1625
NH +3 beds (50% from RG)	569	1060	1629
RG +3 beds (0% from NH)	1060	602	**1662**
RG +3 beds (10% from NH)	1098	544	1642
RG +3 beds (20% from NH)	1135	484	1619
RG +3 beds (50% from NH)	1246	279	1525
NH +4 beds (0% from RG)	972	600	1572
NH +4 beds (10% from RG)	899	729	1628
NH +4 beds (20% from RG)	823	827	1650
NH +4 beds (50% from RG)	564	1080	1644
RG +4 bed (0% from NH)	1077	566	1643
RG +4 beds (10% from NH)	1114	509	1623
RG +4 beds (20% from NH)	1150	451	1601
RG +4 beds (50% from NH)	1252	256	1508

After reviewing the throughput of individual hospitals, it appeared that it is not only increased as a result of a higher percentage of patients transferred, but also as more beds are transferred. For example, throughput at RG for scenarios where 20% of patients are transferred from NH to RG starts from 1107 for one extra bed, 1134 for two extra beds, 1135 for three extra beds and finally increases to 1150 for 4 extra beds.

If the decision regarding the number of beds and the proportion of patients transferred was purely based on the highest overall throughput, NH demand would remain the same with the bed capacity reduced by three, resulting in a very high degree of rejection. As before, it is crucial to investigate variation in the probability of all beds being occupied for each of the 32 scenarios. Table 6.6 presents the probabilities of rejection at individual hospitals and the overall probabilities.

The total probability of all beds being occupied in the base model is equal to 16.69%. A lower probability is obtained for eight scenarios; those are highlighted in red in Table 6.6. The lowest total probability of rejection of 14.04% is observed for the scenario where two beds and no patients are transferred from RG to NH. The highest probability of rejection (44.94%) is experienced when four beds and no patients are transferred from NH to RG.

The probability of rejection is now examined for individual hospitals. Unsurprisingly, as the smaller hospital is transferring more beds to the bigger hospital, the probability of rejection increases very fast in the smaller one. For example, if 10% of patients are transferred from NH to RG together with one bed, the probability is 11.62%; for two beds 25.27%; for three 30.27% and for four it is 39.65%.

The main aim of this section is to provide recommendations to CCU managers regarding the proportion of patients and beds that should be transferred in order to increase overall throughput and decrease probability of all beds being occupied. Table 6.5 and 6.6 are used to draw these conclusions. It appears that there exists only one scenario that satisfies both conditions; to transfer 20% of RG demand to NH along with three beds. As a result, the overall reduced arrival rate at RG is 2.25, which is close to the new increased arrival rate at NH of 2.20. Not only are the arrival rates similar, but also the bed capacities; 11, 12 and 13 in RG in the corresponding periods and 11 beds in NH. In the recommended scenario throughput is increased by 0.87% and probability of rejection decreased by 10.45%.

To conclude, the overall throughput is observed to be higher if the smaller hospital receives extra beds from the bigger hospital. In addition, the overall probability of rejection is lower if the smaller hospital gets extra beds and extra patients from the bigger hospital.

Table 6.6: Probability of rejection for each of the 32 scenarios

Scenario	**Probability of rejection, %**		
	RG	**NH**	**Overall**
Base model	6.42	10.28	**16.69**
NH +1 bed (0% from RG)	8.06	6.83	14.89
NH +1 bed (10% from RG)	5.98	11.25	17.23
NH +1 bed (20% from RG)	4.11	16.11	20.22
NH +1 bed (50% from RG)	0.52	30.54	31.06
RG +1 bed (0% from NH)	5.07	14.85	19.92
RG +1 bed (10% from NH)	6.61	11.62	18.23
RG +1 bed (20% from NH)	8.33	8.49	16.82
RG +1 bed (50% from NH)	14.20	1.51	15.71
NH +2 beds (0% from RG)	10.06	3.98	**14.04**
NH +2 beds (10% from RG)	7.68	7.55	15.23
NH +2 beds (20% from RG)	5.48	11.72	17.20
NH +2 beds (50% from RG)	0.88	25.19	26.06
RG +2 beds (0% from NH)	3.97	25.27	29.24
RG +2 beds (10% from NH)	5.28	21.16	26.43
RG +2 beds (20% from NH)	6.75	16.81	23.56
RG +2 beds (50% from NH)	11.99	4.44	16.43
NH +3 beds (0% from RG)	13.47	1.72	15.19
NH +3 beds (10% from RG)	10.72	3.89	14.61
NH +3 beds (20% from RG)	8.06	6.89	14.95
NH +3 beds (50% from RG)	1.71	18.80	20.51
RG +3 beds (0% from NH)	2.74	34.25	36.99
RG +3 beds (10% from NH)	3.77	30.27	34.04
RG +3 beds (20% from NH)	4.98	25.91	30.88
RG +3 beds (50% from NH)	9.51	11.10	20.61
NH +4 beds (0% from RG)	21.76	0.97	22.73
NH +4 beds (10% from RG)	18.29	2.41	20.70
NH +4 beds (20% from RG)	14.67	4.61	19.28
NH +4 beds (50% from RG)	4.24	14.70	18.93
RG +4 beds (0% from NH)	1.40	43.54	44.94
RG +4 beds (10% from NH)	2.06	39.65	41.71
RG +4 beds (20% from NH)	2.89	35.25	38.14
RG +4 beds (50% from NH)	6.35	18.68	25.03

It has been shown that sharing resources between hospitals is beneficial for both hospitals. The next section will investigate whether combining two Units together to form one big Unit will also be beneficial.

6.4.3 Consolidation of Two Units

During this project new plans for a Specialist Critical Care Centre (SCCC) to be built in Llanfrechfa near Cwmbran in South Wales came up. The SCCC would take services from Royal Gwent Hospital in Newport and Nevill Hall Hospital in Abergavenny. The purpose of this 'what if' scenario is to examine the impact of consolidating two CCUs into a single unit. The first main advantage of the consolidation would be increased bed flexibility. Centralising some services, such as critical care, in fewer large hospitals would allow patients to get access to the best care around the clock.

In order to check whether centralising critical care would prove beneficial to hospital managers and patients, measures such as: bed occupancy variability and the probability of having all beds full will be tested. Before these measures are examined, both data sets have to be merged into one. The combined data set, which includes information regarding 4098 patients, will be analysed in order to understand the current combined CCU system.

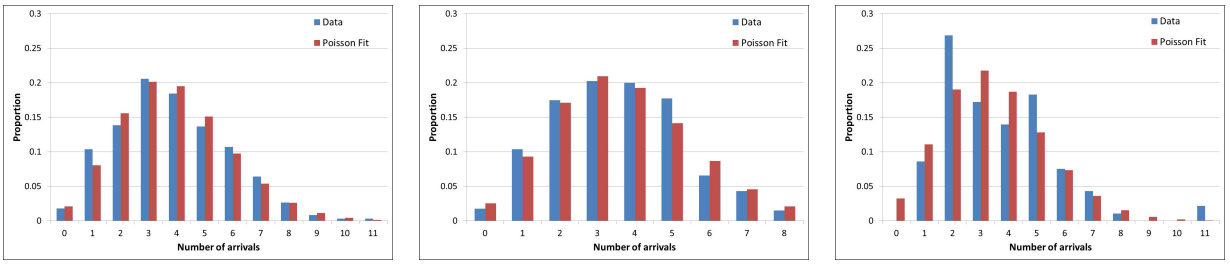
To determine the most appropriate statistical distribution to accurately represent the profile of arrivals in the queueing model, initially the nature of admissions to the combined CCU will be explored. The number of arrivals on each day in the combined CCU is the sum of the number of admissions in individual CCUs. Again, extra attention has to be put towards the fact that bed capacity changed twice. In Period 1 there were 22 beds in total, in Period 2 there were 23 beds and in the final period 24 beds. Summary statistics for the daily number of arrivals at the CCU separately for each of the three periods and the overall are given in Table 6.7.

Table 6.7: Summary statistics for the number of admissions in the combined hospital

Summary statistic	**Period 1**	**Period 2**	**Period 3**	**Overall**
Mean	3.867	3.565	3.688	3.742
Median	4	4	3	4
Standard Deviation	2.006	1.734	2.016	1.917
Minimum	0	0	1	0
Maximum	11	8	11	11

The overall average admission rate is 3.742 patients per day and there is some variation across the three periods; therefore, the three frequency distributions will be considered separately. *@Risk* is utilised to determine the appropriate distributions; it is suggested that a Poisson distribution

provides the best fit. The determined parameter values of λ in Period 1 is 3.874, in Period 2 it is 3.676 and 3.432 in Period 3; correspondingly giving the value of the sum of squared differences between the data and the analytical results of 0.0014, 0.0020 and 0.0157. Consider the frequency distributions of the number of arrivals per day during each period, presented in Figures 6.8a, 6.8b and 6.8c.



(a) Period 1 (b) Period 2 (c) Period 3

Figure 6.8: Frequency distribution of the number of arrivals for each period

Figures 6.8a and 6.8b demonstrate the goodness of the Poisson fit. Figure 6.8c does not provide a very good fit, as a result of very small amount of data (only 343 people admitted in Period 3 that lasted 92 days).

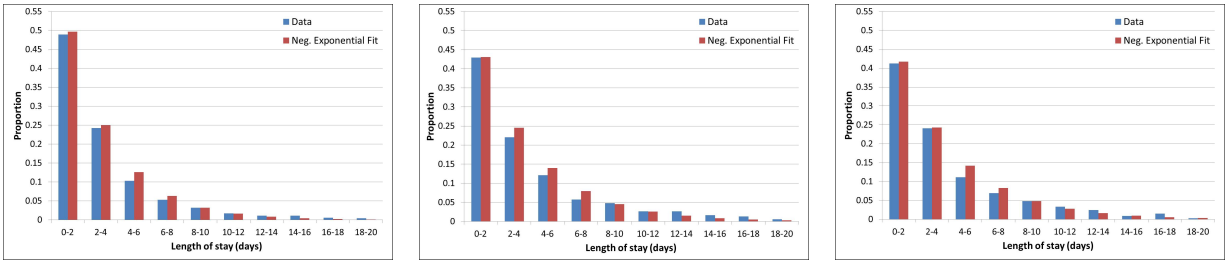
The second measure that influences the flow of patients through the CCU is length of stay. The combined data set is analysed and a number of relevant summary statistics is considered; these are presented in Table 6.8.

Table 6.8: Summary statistics for length of stay in the combined hospital

Summary statistic	**Period 1**	**Period 2**	**Period 3**	**Overall**
Mean	4.186	5.103	4.846	4.556
Median	2.042	2.6	2.8	2.193
Standard Deviation	8.147	8.385	6.951	8.150
Minimum	0.014	0.014	0.014	0.014
Maximum	193.094	152.4	61	193.094

The overall average LoS is 4.56 days and there is a significant variation across the three periods, therefore three frequency distributions will be considered separately. As before the statistical tool *@Risk* is utilised to determine the appropriate distributions; it is suggested that a Negative Exponential distribution provides the most accurate fit. The determined parameter values of μ in Period 1 is 0.344, in Period 2 is 0.281 and 0.270 in Period 3; correspondingly giving the value of the sum of squared difference between the data and analytical results of 0.0009, 0.0018 and 0.0015. Consider the frequencies displaying LoS grouped in two day classes for each period, presented in

Figures 6.9a, 6.9b and 6.9c. 2.51% of observations in Period 1, 2.47% in Period 2 and 2.71% in Period 3 (LoS greater than 20 days) are excluded form the graphs for presentation purposes. The Negative Exponential distribution clearly provides a good fit to LoS.



(a) Period 1 (b) Period 2 (c) Period 3

Figure 6.9: Length of stay frequencies for each period

Finally, the number of patients present in the combined Unit during every hour in the period from 01/04/2009 to 31/12/2011 is obtained. The number of beds occupied at every hour in the study period is illustrated in Figure 6.10; the periods are separated by the two vertical lines.

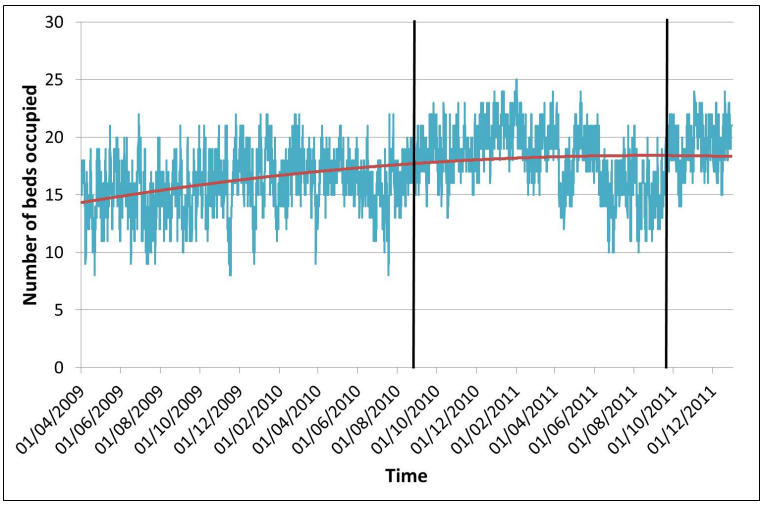


Figure 6.10: Bed occupancy at each hour of the study period

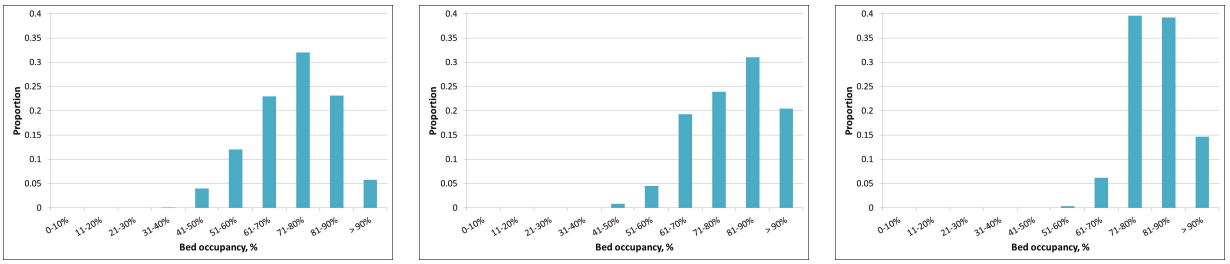
Summary statistics of bed occupancy in each period are presented in Table 6.9.

Table 6.9: Bed occupancy summary statistics in the combined hospital

Summary statistic	**Period 1**	**Period 2**	**Period 3**	**Overall**
Mean	16	18.274	19.536	17.221
Median	16	19	20	17
Standard Deviation	2.433	2.667	1.916	2.8072
Minimum	8	10	14	8
Maximum	22	25	24	25

The overall mean bed occupancy is 17.22; however, the variation across the periods is visible. The increase in bed capacity causes an increase in the average bed occupancy, which is also indicated by the increasing trend-line in Figure 6.10. An initial analysis reveals that the probability of having all beds occupied differs throughout. In Period 1 all beds were full on 0.6% of occasions, in Period 2 on 3.83% and in Period 3 on 0.09%.

As previously, the percentage bed occupancy distribution is considered rather than the typical distribution. Consider the frequencies displaying the bed occupancy percentage distribution for each period, presented in Figures 6.11a, 6.11b and 6.11c.



(a) Period 1 (b) Period 2 (c) Period 3

Figure 6.11: Bed occupancy frequencies for each period

As before in Section 6.3, the three frequency distributions are combined into one by taking the proportion of duration each period lasted (Period 1, 17 months; Period 2, 13 months and Period 3, 3 months). Figure 6.12 illustrates the overall bed utilisation frequency distribution, which will be used in Section 6.5 for the comparison between the data and the model results.

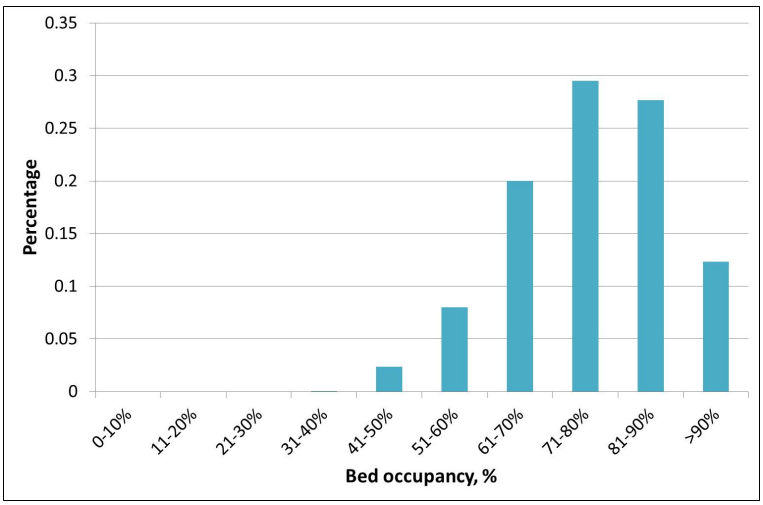


Figure 6.12: Overall bed occupancy frequency distribution for the combined Unit

To decide whether consolidation of the two Units would be beneficial, two variables are examined.

Obviously, low bed occupancy variation makes the system more stable, more predictable and less stressful to the clinicians. The first variable that will be tested is the coefficient of variation (CV) of bed occupancy, which is a normalised measure of dispersion; it is defined as the ratio of the standard deviation to the mean, and is expressed as a percentage. Table 6.10 displays calculated CVs for NH, RG and the combined Unit.

Table 6.10: Comparison of the coefficients of variation

Period	**NH**	**RG**	**Average of two individual Units**	**Combined hospital**
Period 1	26.4%	17.5%	22%	15.2%
Period 2	25.7%	14%	19.9%	14.6%
Period 3	19.2%	11.8%	15.5%	9.8%

The columns labelled 'NH' and 'RG' show how the coefficient of variation changed over the time in each CCU. As expected, the smaller of the two Units experiences higher degree of variation. The 'Average of two individual Units' is an average of the CVs in NH and RG. The final column gives the CV of the combined Unit. Evidently, it is lower in each period, by at least 5.3%.

The second examined variable is the probability of the system being full. The system, where probability of rejection is lower is obviously preferred, not only by patients, but also by the staff. Table 6.11 displays the percentage of occasions when the individual Units and the combined Unit were full.

Table 6.11: Comparison of the probability of having all beds full

Period	**NH**	**RG**	**Combined hospital**
Period 1	6.1%	3.3%	0.6%
Period 2	13.9%	10.4%	3.8%
Period 3	10.6%	3.5%	0.01%

Clearly, in the smaller hospital, NH, the proportion of time that all beds were occupied is higher than in RG during every period. Most importantly, the probability of rejection is significantly lower in the combined Unit, proving the benefits of having one big Unit.

As a result of the consolidation, it can be concluded that the bed occupancy variability is decreased, also the probability of having all beds full is significantly decreased, implying that bed flexibility is increased and the system is more stable and controllable.

It has been established that a combined Unit would be beneficial, confirming the existing plans of the SCCC to take over services at NH and RG. The question arises whether the current bed capacity of 24 beds is sufficient for the new Unit. The next section will investigate how many beds the SCCC should have.

6.5 SCCC Capacity Recommendations

6.5.1 Introduction

Managers in healthcare are frequently faced with decisions on re-dimensioning of services. Changes in capacity are typically based on estimations using mean admission numbers, and mean lengths of stay. However, deciding on how many beds to provide is not a simple task. Mathematical models help to decide how well services perform with given bed capacity. The results however, require a risk judgement. A balance must be struck between risk of cancellation because the Unit is full and the economic benefits of full utilisation of resources. This predicament is becoming more prevalent in the healthcare community as indicated by the significant body of literature that considers the allocation of scarce healthcare resources. Section 6.5.2 reviews work on determining system capacity, based on desired system goals and requirements.

6.5.2 Literature Review

6.5.2.1 Bed Allocation and Planning

The demand for hospital beds can be divided into elective (scheduled) and emergency (unscheduled) admissions. Both categories of admissions impact on how many beds are needed to meet demand, while maintaining reasonable bed utilisation rates. In the literature, most bed planning queueing models attempt to overcome bed shortages or policies that lead to patient misplacement, bumping, or rejection. Hospital managers are under pressure to reduce bed capacity and decrease occupancy rates in the name of operational efficiency.

Young (1962a [172], 1962b [171]) proposed an incremental analysis approach in which the cost of an additional bed is compared with the benefits it generates. Beds were added until the increased cost was equal to the benefits.

Kao *et al*, 1981 [97] proposed an $M/G/\infty$ model to periodically reallocate beds to medical services to minimise the expected overflows in different hospital wards. A demand forecasting system using the existing data base for generating inputs to the bed allocation model was constructed. For each medical service, the authors established the base line requirement by requiring that the number of beds assigned to the service should be sufficient, in terms of yearly aggregate, to accommodate a

pre-specified amount of patient load generated by the target population. The authors distributed the remaining beds to services with the objective of minimising the expected total average overflows over the months in the planning horizon.

Green *et al*, 2001 [70] applied a queueing model approach to the hospital bed planning issue to gain insights on the potential impact of cost-cutting strategies on patients' delays for beds. Using a queueing theory approach, the factors that have the greatest impact on the trade-off between hospital occupancy levels and delays were identified. It was stated that using target occupancy levels as the primary determinant of bed capacity is inadequate and may lead to excessive delays for beds.

By integrating queueing theory and compartmental models of flow, Gorunescu *et al*, 2002 [62] demonstrated how by changing arrival rates, length of stay and bed allocation influences bed occupancy, emptiness and rejection rates in departments of geriatric medicine in a London teaching hospital. By considering an $M/PH/c/n$ queue model, the authors showed how the provision of extra emergency use beds could improve performance while controlling costs.

de Bruin *et al*, 2005 [39] applied a stationary 2-D queueing system with blocking to analyse congestion in emergency care chains. The primary goal was to determine the optimal bed allocation over the emergency care chain, given a required service level (maximum of 5% refused admissions). The bottlenecks were identified, the impact of fluctuation in demand was described and the optimal bed capacity distribution for cardiac patients was calculated. Cooper *et al*, 1974 [34] dealt with a very similar problem extended to estimating the number of beds necessary for two units: acute and intermediate coronary care, each of which should have a maximum turn-away rate of 5%.

Koizumi *et al*, 2005 [104] analysed congestion levels in the Philadelphia mental health system using a queueing network model with blocking. Their model focused on blocking between three types of mental facilities, namely, extended acute hospitals, residential facilities and supported housing. The authors investigated how effectively the increase in the number of supported housing beds could reduce the steady-state congestion level in the system.

Cochran *et al*, 2006 [31] proposed a multi-stage stochastic methodology for analysing the flow of patients in a whole hospital setting with multiple patient types. The authors combined queueing network analysis and discrete event simulation to balance bed utilisation targets with the associated benefits of reduced: waiting, patient blocking, poor bed assignment and emergency department overflow behaviours. The methodology is applied to a 400 bed major hospital including an emergency unit.

Cochran and Bharti, 2006 [31] proposed a four-stage stochastic methodology for bed planning in

hospitals using queueing networks and discrete event simulation models. The methodology aims to balance bed unit utilisations in an entire hospital and minimise the blocking of beds from upstream units within given constraints on bed reallocation. The methodology included the assessment and effect of time-dependent patterns. Queueing networks were used to assess the flows between units and to establish target utilisations of bed units. Discrete event simulation was then used to maximise the flow through the balanced system including non-homogeneous effects, non-exponential lengths of stay, and blocking behaviour.

Chaussalet *et al*, 2006 [26] developed a patient flow model through healthcare systems with constrained capacity. The model used a closed queueing network with the assumption that the system is always full. The authors modelled the progression of patients through a geriatric department in the UK as a set of conceptual phases. On admission, patients enter the first phase (assessment, diagnosis, etc.), from which they are either discharged, or transferred in to the second phase (some form of rehabilitation). In the final phase, which corresponds to long-stay care, all patients are eventually discharged. The model assisted service managers and clinicians with decision-making on bed allocation and on discharge policies.

de Bruin *et al*, 2007 [40] investigated the emergency in-patient flow of cardiac patients in a university medical centre. The impact of variability (in both length of stay and arrivals) on capacity requirements was described. They applied a queueing model to analyse congestion in the emergency care chain. With this model, the number of beds in the care chain is determined for several service levels, given a maximum number of refused admissions. In 2009, de Bruin *et al* [38] developed a decision support system, based on the Erlang loss model, which can be used to evaluate the current size of nursing units, and to quantify the impact of bed reallocations and merging of wards.

6.5.2.2 Bed Allocation in Critical Care Units

Ridge *et al*, 1998 [138] presented a simulation model of an $M/M/c$ queue that demonstrated methods for capacity planning in a six-bed ICU. They showed that there is strong non-linear correlation between the number of beds available, mean occupancy level and the number of transfers due to lack of free beds. The model did not allow admission of elective patients when the majority of beds were occupied. The effect of a rudimentary deferral rule for elective patients was investigated.

Green, 2003 [64] examined data from New York State and used queueing analysis to estimate bed unavailability in ICUs and obstetrics units. Using various patient delay standards, units that appear to have insufficient capacity are identified. The results indicate that as many as 40% of all obstetrics units and 90% of ICUs have insufficient capacity to provide an appropriate bed when needed. This

contrasts sharply with what would be deduced using standard average occupancy targets. According to the authors, smaller ICUs should operate at much lower occupancy percentages. It was also shown that by doubling admission rates, the standard deviation of the number of admissions is not doubled and hence the hospital does not need to double the number of beds.

Shmueli *et al*, 2003 [149] presented a queueing theory model for optimising admissions to an ICU, where the objective is to maximise the expected incremental number of lives saved from operating the ICU. A single-queue models to find the probability distribution of the number of occupied ICU beds were used. The authors modelled the ICU at Jerusalem Hebrew University-Hadassah Hospital by using the proposed methodology and showed that a relative life saving improvement of 17.9% could be achieved by reforming the ICU admission policy. Three different policies were considered: first-come-first-served, first-come-first-served for all referrals whose expected incremental survival benefits gained from ICU admission exceed some hurdle, and first-come-first-served for all referrals whose expected incremental survival benefits exceed a bed specific hurdle that depends upon the number of occupied beds.

McManus *et al*, 2004 [122] demonstrated a queueing model, which accurately predicts the bed occupancies in a busy ICU. The predictions from the model to the data were compared and the sensitivity of the model to changes in the number of beds available was explored. The model was useful in predicting both monthly responsiveness to changing demand and the overall two year turn-away rate for the unit. It was showed that bed availability is dependant on staffing shortages or admission of patients with a very long length of stay. A significant correlation between occupancy and turn-away rates was presented.

Griffiths *et al*, 2005 [76] proposed a $M/H/c/\infty$ queueing model of the ICU environment, with particular emphasis on adequately representing the high variation in the patient's length of stay. It was anticipated that the model will be utilised as a tool for resource management. In conjunction with the queueing model, a discrete-event simulation model enabling patients to be classified according to source of admission was developed.

Cochran *et al*, 2007 [32] stated that financial data should be used and not census data for estimating inpatient bed capacity, since it estimates true demand for service rather than the service available to be offered. The queueing theory model was used to estimate the number of beds at a level one trauma facility in the United States.

Seshaiah *et al*, 2011 [147] studied the patient flows to and between the different units of a hospital. A queueing network model with blocking and reneging was developed to study how the number of available beds in ICU and General Units influence the wait times in the Emergency Care Units.

The authors studied the $M/G/k$ queue with reneging and through approximate logical methods and simulation the adequate bed counts in each of the two units was determined so as to guarantee certain access standards.

The significant body of literature considered the changes in capacity, the next section will add to this by informing as to sufficient number of beds required in the combined Unit.

6.5.3 Bed Occupancy Model

Mathematical models can help to decide how well services perform with a given bed capacity. The results however, require a risk judgement. Assuming that no hospital is able to fund enough beds to cover all demand peaks, a decision must be made about how often it is acceptable to allow the Unit to be full.

Before any changes to the Unit size are explored, the bed occupancy model that accurately represent the consolidated Unit must be obtained. Having accurate arrival and length of stay fitted distributions, described in detail in Section 6.4.3, allow to model the combined Unit using the analytical results described in Section 6.2.

The number of beds available is chosen to be 23, since it is the rounded product of the number of beds available in each period and proportion of time each period lasted. The overall service rate is calculated in the following way:

$$\mu = P(\text{Period 1}) \times \mu_{\text{Period 1}} + P(\text{Period 2}) \times \mu_{\text{Period 2}} + P(\text{Period 3}) \times \mu_{\text{Period 3}} = 0.219$$

The next task is to determine values for the parameters $k_1, k_2, \lambda_a, \lambda_b, \lambda_c$ which would give the closest representation of bed occupancy probabilities illustrated in Figure 6.12. *Evolver* is utilised to find these parameters to match the mathematical model with bed occupancy probabilities, which are grouped in ten classes: $0-10\%, 10-20\%, 20-30\%$ etc. In Section 6.2 *Evolver* was utilised to minimise the sum of squared differences between $P_{\text{data}}(n)$ and $P_{\text{model}}(n)$ for $n = 0, \ldots, c$, where n is the number of beds occupied, for each period. The mathematical program used to find model parameters is described below:

Minimise:

$$\sum_{i=1}^{c} \left(P_{\text{data}}(10s\% < n \leq 10(s+1)\%) - P_{\text{model}}(10s\% < n \leq 10(s+1)\%)\right)^2$$

Subject to:

$$0 < k_1 < k_2 < c$$

$$k_1, k_2 \in \mathbb{N}$$

$$\lambda_a, \lambda_b, \lambda_c \in \mathbb{R}_+$$

$$0.9\lambda_{\text{data}} \leq \Lambda \leq 1.1\lambda_{\text{data}}$$

where:

$$\lambda_{\text{data}} = P(\text{Period 1}) \times \lambda_{\text{Period 1}} + P(\text{Period 2}) \times \lambda_{\text{Period 2}} + P(\text{Period 3}) \times \lambda_{\text{Period 3}} = 3.742$$

$$\Lambda = \sum_{n=1}^{k_1} \lambda_a P_n + \sum_{n=k_1+1}^{k_2} \lambda_b P_n + \sum_{n=k_2+1}^{c} \lambda_c P_n$$

$$s \in [0, \ldots, 9]$$

The mathematical program is solved heuristically using a genetic algorithm (implemented with *Evolver*) and the obtained parameters are given in Table 6.12.

Table 6.12: The parameter values for the combined hospital model

Parameter	**Parameter value**
λ_a	3.675
λ_b	4.746
λ_c	2.746
k_1	14
k_2	19

The analytical results obtained for the set of parameters are now compared with the data in Figure 6.13. Visually, there is close agreement in bed occupancy levels. Other factors suggesting that the model provides a good fit are listed in Table 6.13.

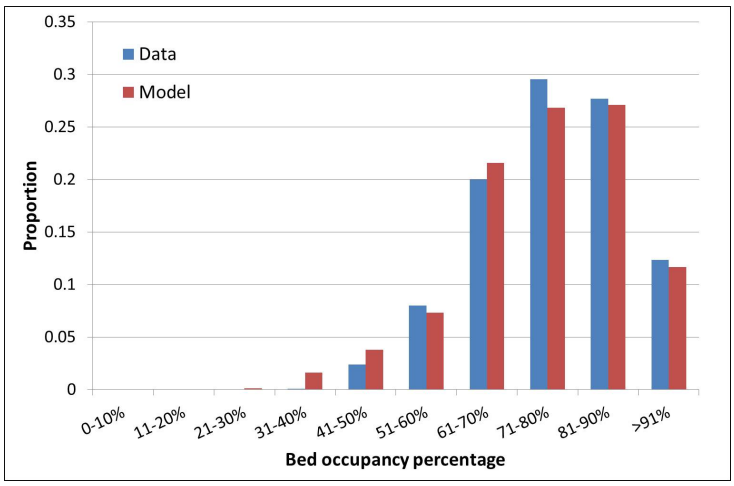


Figure 6.13: Comparison of bed occupancy model and data

Table 6.13: Model validation

Variable	**Model**	**Data**
Overall arrival rate, Λ	4.114	3.742
Bed occupancy utilisation	74.90%	76.16%
Probability of all beds full	1.92%	1.83%
Probability of $> 80\%$ of beds occupied	38.75%	40.04%

The model variable values compares favourably with the observed values, therefore it can be concluded that the model satisfactorily represents the real world system, and now it can be tested to changes in Unit size.

6.5.4 Changes in the Unit Capacity

A decision regarding the number of beds in the SCCC will be based on how often it is acceptable to allow the Unit to be full and how many beds on average should be occupied. The managers from the Aneurin Bevan Health Board indicated that the Unit should be full on less than 5% of occasions. Although the 85% bed occupancy target is the most often quoted in the literature (Green, 2003 [64]), the managers decided on 80% bed occupancy utilisation rate. Having these two targets in mind, the base model with 23 beds will be adapted to allow bed capacity to be changed. It is decided to test how performance of the Unit changes assuming there are $19, 20, \ldots, 30$ beds available. The arrival rates $\lambda_a, \lambda_b, \lambda_c$ and service rate, μ, remain unchanged. The cut-off points must be proportionally adjusted in the following way: $\widehat{k_1} = \left\lfloor \frac{\text{original } k_1}{\text{original bed capacity}} \times \text{new bed capacity} \right\rceil$ and similarly $\widehat{k_2} = \left\lfloor \frac{\text{original } k_2}{\text{original bed capacity}} \times \text{new bed capacity} \right\rceil$. For each bed capacity the collected measures are:

- the probability of rejection, $P(c)$;
- the utilisation rate, $\frac{\sum_{n=1}^{c} nP(n)}{c}$;
- the probability of having more than 80% of beds occupied, $P(n \geq 80\%) = \sum_{n=\lceil 80\% c \rceil}^{c} P(n)$;
- the probability of having more than 90% of beds occupied, $P(n \geq 90\%) = \sum_{n=\lceil 90\% c \rceil}^{c} P(n)$.

Figure 6.14 illustrates how the measures of interest change as a result of an amendment to the number of beds.

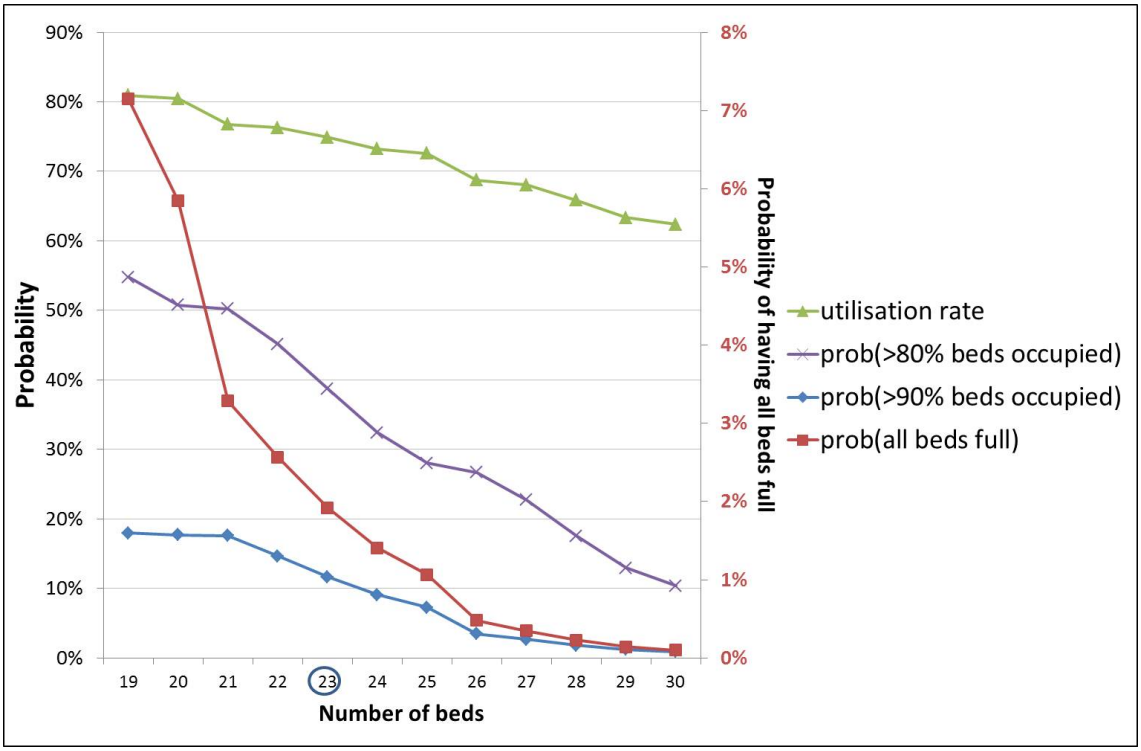


Figure 6.14: Effect of bed capacity change with the base model of 23 beds

The green line, with the scale on the left hand side, shows how the bed utilisation rate is affected by the different number of beds available. The purple and blue lines, with the scale also on the left hand side, illustrate the change in the probability of having more than 80% and 90% of beds full respectively. The red line, with the scale on the right hand side, shows the change in the probability of rejection.

The utilisation rate fell below 80% and similarly, the probability of all beds being full fell below 5% for a bed capacity of 21. This suggests that, if the two Units were to be combined, the current capacity of 23 would be sufficient to satisfy the two targets, again confirming what was concluded in Section 6.4.3 that having one big Unit is more beneficial than two smaller ones. However, the purple line (the probability of having more than 80% of beds occupied) suggest that on more than 50% of occasions the Unit is running on a relatively high bed occupancy, which might prove stressful to the clinicians. If the percentage of time that the Unit is busy was decreased from 50% to 30%, then the suggested number of beds would be 25. That would result in the probability of rejection of 1.07%, the utilisation rate of 72.57% and the probability of having more than 80% of beds full on 28% of the time.

In conclusion, the scenario proposed helps to inform the decision making regarding the most suitable number of beds in the combined Unit. The next two sections will investigate whether the bed

capacity recommendation changes by altering the arrival rate or service rate.

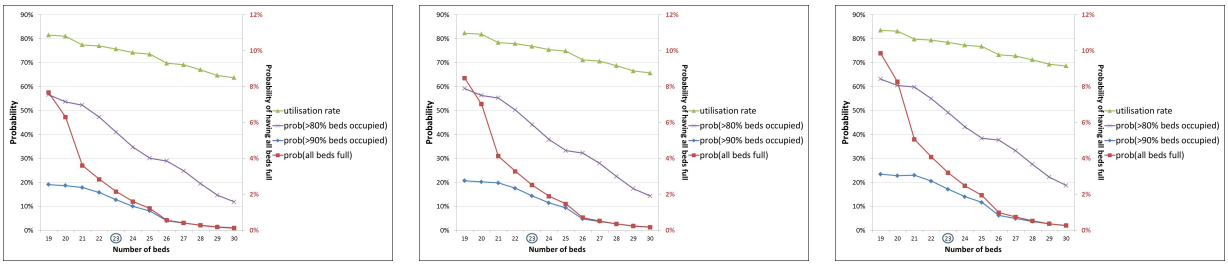
6.5.4.1 'What if' Scenario: Increased Arrivals

Having a brand new hospital might increase attractiveness and more patients might decide to choose to be admitted there. The second aspect, that might potentially increase the number of patients arriving at the SCCC is increased area coverage and also an ageing population (Office for National Statistics, 2012 [132]). Also, the data sets provided did not include information regarding patients that were admitted to an ordinary ward within the hospital, or were transferred to different CCU, due to the lack of spare beds in the CCU. This section will investigate the impact on bed requirements by increasing the arrival rate by 2%, 5% and 10%.

The arrival parameters are changed in the following way:

$$\lambda_a^{new} = \lambda_a \times (1 + x\%), \lambda_b^{new} = \lambda_b \times (1 + x\%) \text{ and } \lambda_c^{new} = \lambda_c \times (1 + x\%)$$

where x is the increase percentage. The same measures as previously are calculated, and Figures 6.15a, 6.15b and 6.15c demonstrate how the different measures are affected by the bed capacity.



(a) 2% increase (b) 5% increase (c) 10% increase

Figure 6.15: Effect of bed capacity change with the corresponding admission rate increase

If the decision regarding the number of beds was only based on 80% bed utilisation and 5% rejection rate targets, for the 2%, 5% and 10% arrival increase rate the recommended number of beds would be 21, 21 and 22 respectively. If the constraint is added regarding having a busy Unit with more than 80% of beds occupied on less than 30% of the time, then the suggested bed capacities would be 26, 27 and 28 respectively. Therefore, having a small increase in the overall admission rate implies more beds are needed to run the Unit smoothly.

6.5.4.2 'What if' Scenario: Bed Blocking Reduction

The human decisions regarding admissions and discharges are based on the CCU bed occupancy levels. Medical staff may therefore slightly delay discharge or authorise slightly earlier discharge if the patient's condition allows. The main outcome of interest is to investigate how varying the

mean delay to discharge impacts the system. Green and Nguyen, 2001 [70] stated that reducing average LoS has far more potential to reduce required capacity than reducing LoS variability. The most sensible option for reducing average LoS is to deal with the problem of delayed discharge. A number of authors have studied the impact of bed blocking on the number of hospital beds desired.

In systems with blocking, congestion not only increases patient waiting time, but also reduces the throughput of the system. de Bruin *et al.*, 2005 [39] determined the number of beds required to achieve a maximum turn away rate of 5% at the emergency cardiac department of the university medical centre of Amsterdam, which implements the pure loss model. Cooper and Corcoran, 1974 [34] dealt with the same problem extended to a sequence of two stations each of which should have a maximum turn-away rate of 5%. Milliken *et al.*, 1972 [125] sought a 1% turn-away rate in an obstetrics department in which vaginal births have priority over scheduled caesarian sections. The authors pointed out the benefits of economies of scale, so that larger facilities incur lower bed investment per additional birth.

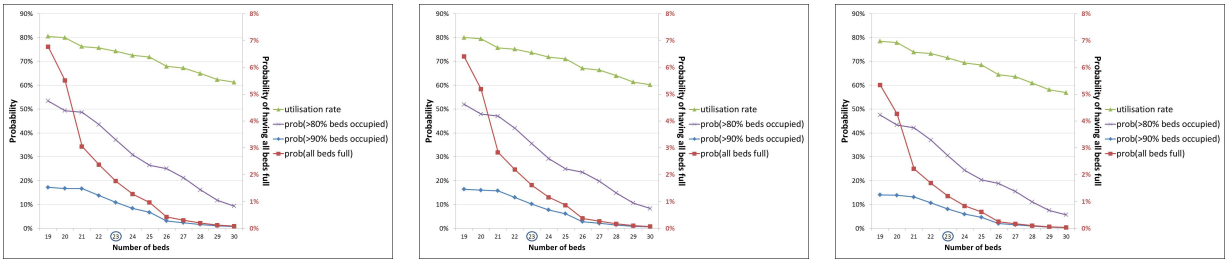
Given a desired maximum turn-away rate, de Bruin *et al.*, 2007 [40] determined the optimal number of beds in a cardiology department. The cardiology department was modelled as a network of three sub-departments. The research found that too few beds downstream is the primary cause of refused admissions upstream and that congestion effects can add 20-30% to patient length of stay in the department. It was claimed that having a fixed target utilisation rate is unrealistic and concluded that a downstream utilisation of 55% is necessary to attain a 2% turn-away rate. As an alternative, departments could be merged to gain the benefits of economies of scale thereby meeting the goal at higher occupancy rates.

Blair and Lawrence, 1981 [15] sought to design the capacity of horizontally integrated burn care facilities throughout the state of New York, so that no more than 5% of patients are turned away from the system. If a patient goes to a facility which is fully occupied, that facility would refer to the patient to another which is not filled. If all facilities are fully occupied, the patient is lost to the system. Queueing theory was used to determine the capacity of the entire system as if it were one queueing system. This capacity is then allocated to facilities in a manner that best attains their individual goals. They find such a system-planned approach ideal for a system with low demand and high infrastructural costs.

In the study period of the joined hospital there were 67.7% of patients who experienced any delay in their discharge. Patients' delay ranged from 10 minutes to 21 days (mean 15.89 hours; standard deviation of 1.29 days; median 4 hours).

Patient LoS can be divided into two phases: active treatment phase and bed blocking phase, hence

the overall mean service rate is as following: $\mu = \mu_{\text{active}} + \mu_{\text{blocked}}$. The hospital does not have much control over the active treatment time, μ_{active}; however, bed blocking time can be reduced. The effect of decreasing the average delay to discharge time by 10%, 20% and 50% will be inspected. The overall mean service rate is $\mu = \mu_{\text{active}} + \mu_{\text{blocked}} \times (1 - x\%)$.



(a) 10% decrease (b) 20% decrease (c) 50% decrease

Figure 6.16: Effect of reducing bed blocking by the corresponding percentage

If the decision regarding the sufficient number of beds was only based on 80% bed utilisation and 5% rejection rate, for the 10%, 20% and 50%, bed blocking reduction rate the recommended number of beds would be 21, 21 and 20 respectively. If the constraint is added regarding having a busy Unit with more than 80% of beds occupied less than 30% of the time, the suggested bed capacities would be 25, 24 and 24 respectively. Reducing by 10% does not change the suggested bed capacity; however, reducing the average mean blocking time by 20% or 50% reduces the recommended bed capacity by one. If the mean delayed discharge was eradicated, the CCU would require only 22 beds to meet both targets and the guideline regarding having more than 80% of beds occupied on less than 30% of occasions.

6.6 Conclusions

It has been shown that a mathematical model which omits the fact of different arrival rates for different bed occupancy levels did not adequately describe the CCU activities at the Royal Gwent and the Nevill Hall CCUs. The influence of management policies for patient admission depends on the bed occupancy levels. This chapter has focussed on developing a mathematical model of patients admitted under the care of Aneurin Bevan Health Board, which included the dependency of admission rate on actual occupancy.

Using the developed model, the potential improvement of operational efficiency by transferring beds and patients from one Unit to other has been shown. The scenario of merging CCUs and the benefits from it have been illustrated.

The work described in this chapter has helped to inform capacity planning decisions at the SCCC.

A few 'what if' scenarios were investigated. The model has been used to explore a scenario, which will investigate the effect of increasing the arrival rate. The issue of bed blocking has been documented and effect of decreasing the delay time has been investigated.

Chapter 7

A Game Theoretical Consideration of Critical Care Unit Interaction

7.1 Motivation of the Study

The work described in Section 6.4.3, where two Critical Care Units (CCUs) were combined into one, proved that cooperation between CCUs is beneficial. It has also been shown in Chapters 5 and 6 that there are noticeable behavioural aspects apparent; for example, leaving patients for longer in the Unit if there is no pressure on CCU beds, or admitting less patients if bed occupancy levels are high. In this part of the thesis non-cooperative game theoretical models of two CCUs will be proposed, where each CCU acts selfishly, to study the impact of lack of collaboration.

7.2 Literature Review

Game theory is a study of mathematical models of conflict and cooperation between intelligent rational decision-makers. A game is a description of strategic interactions that includes the constraints on the actions that the players can take and the players' interests.

The basic concepts of game theory have been discussed by a number of authors: a book by von Neumann and Morgenstern, 1944 [164] is quoted as the first one, an early textbook by Luce and Raiffa, 1957 [115] or a book by Schelling, 1960 [146] provide discussions of some of the main ideas of the theory. Aumann, 1985 [8] contains a discussion of the aims and achievements of game theory, and Aumann, 1987 [9] is an account of game theory from a historical perspective. Book by Binmore, 1987/88 [14] is a critical discussion of game theory that makes the distinction between the steady state and deductive interpretations. Kreps, 1990 [110] is a reflective discussion of many issues in game theory. Roughgarden, 2005 [144] studied the loss of social welfare caused by selfish, uncoordinated behaviour in networks, and discussed several methods for improving the price

of anarchy with centralized control.

In this chapter and the next, the type of game that will be considered will be referred to as a normal form game. A **normal form game** is a model of a situation in which each player chooses his plan of action once and for all, and all players' decisions are made simultaneously (that is, when choosing a plan of action each player is not informed of the plan of action chosen by any other player). It is assumed that each decision-maker is 'rational' in the sense that they are aware of their own alternatives and also those of their opponents, form expectations about any unknowns, have clear preferences, and choose their action deliberately after a process of optimization. A normal form game consists of:

- a finite set N (the set of players),
- for each player $i \in N$ a non-empty set A_i (the set of actions available to player i),
- each player's preferences are specified by giving a utility function $U : C \rightarrow \mathbb{R}$, which defines a preference of x over y by $U(x) \geq U(y)$.

Some popular normal form games with just two players and each player having only two possible actions include "Battle of the Sexes" (Luce and Raiffa, 1957 [115]) and/or "The Prisoner's Dilemma" (which first entered the literature in unpublished papers by Raiffa in 1951 and Flood in 1952). The game is presented as follows: two criminals are arrested and imprisoned, each prisoner is in solitary confinement with no means of speaking to the other. The police do not have enough evidence to convict the pair on the principal charge and simultaneously question each prisoner. The prisoners can choose between two moves, either "cooperate" or "defect". If one prisoner defects against the other prisoner, he will go free while the other will get three years in prison. If both prisoners defect against each other, both will be sentenced to two years in jail, while if both cooperate, they will be sentenced to one year in jail. The prisoner's dilemma is a canonical example of a game that shows why two players might not cooperate, even if it appears that it is in their common best interest to do so. Each cell of the matrix (Table 7.1) shows the pay-offs in a prisoner's dilemma.

Table 7.1: The normal form game for the prisoner's dilemma

	Prisoner B stays silent (cooperates)	**Prisoner B betrays (defects)**
Prisoner A stays silent (cooperates)	(1,1)	(3,0)
Prisoner A betrays (defects)	(0,3)	(2,2)

Here, regardless of what the other decides, each prisoner gets a lower sentence by betraying the other; however, if they both stayed silent each would receive a lower sentence.

Nash equilibrium is one of the most basic concepts in game theory. The notion of Nash equilibrium was formalised in the context of an abstract strategic game by Nash, 1950 [127], however the basic idea goes back to Cournot, 1838 [36]. Let a_i be a strategy profile of player i and a_{-i} be a strategy profile of all players except for player i. A strategy profile $\hat{a} \in A$ is a Nash equilibrium ([133]) of a strategic game with the property that for every player $i \in N$:

$$U(\widehat{a_{-i}}, \widehat{a_i}) \geq U(\widehat{a_{-i}}, a_i) \text{ for all } a_i \in A_i$$

Thus for $\widehat{a}$ to be a Nash equilibrium it must be that no player i has an action yielding an outcome that he prefers to that generated when they choose $\widehat{a_i}$, given that every other player j chooses his equilibrium action $\widehat{a_j}$. Briefly, no player can profitably deviate, given the actions of the other players. In other words: for any given set of opponent's strategies a_{-i}, where $a_{-i} \in A_{-i}$ define $B_i(a_{-i})$ to be the set of player i's best actions given a_{-i}:

$$B_i(a_{-i}) = \{a_i \in A_i : U(a_{-i}, a_i) \geq U(a_{-i}, a_i^{'}) \text{ for all } a_i^{'} \in A_i\}.$$

I.e., any action in $B_i(a_{-i})$ is at least as good for player i as every other action of player i, $a_i^{'}$, when the other players' actions are given by a_{-i}. The set-valued function B_i is called **the best response function** of player i. A Nash equilibrium is a profile $\widehat{a}$ of actions for which

$$\widehat{a_i} \in B_i(\widehat{a_{-i}}) \text{ for all } i \in N.$$

Another popular game is that of matching pennies. It is played between two players; each player has a penny and must turn the penny to heads or tails. If the pennies match, Player 1 keeps both pennies, if the pennies do not match, Player 2 keeps both pennies. The game can be written in a pay-off matrix (Figure 7.1). Each cell of the matrix shows the two players' pay-offs.

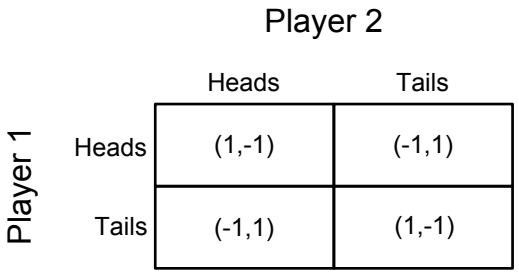


Figure 7.1: Pay-offs in the matching pennies games

This game has no pure strategy Nash equilibrium since there is no pure strategy that is a best response to a best response. Instead, the unique Nash equilibrium of this game is in mixed strategies: each player chooses heads or tails with equal probability. In this way, each player makes the other indifferent between choosing heads or tails, so neither player has an incentive to try another strategy.

The work included in this thesis concentrates only on pure strategy Nash equilibria, i.e. the players' actions are deterministic. The set of pure strategy Nash equilibria of a normal form game is a subset of its set of mixed strategy Nash equilibria. The notion of mixed strategy Nash equilibrium is designed to model a steady state of a game in which the participants' choices are not deterministic but are regulated by probabilistic rules. Mixed equilibria are inadequate in modelling CCUs; the bed capacity is considered to be deterministic as it is expensive and is a long term investment, so the number of beds can not be altered from day to day.

The prisoner's dilemma showed that if both prisoners cooperated, each would get only one year in prison; however, since both prisoners act selfishly, both defected (Nash equilibrium) resulting in two years in prison. A concept in game theory that measures how the efficiency of a system degrades due to selfish behaviour of its agents is called the **Price of Anarchy (PoA)**. The term PoA was first used by Koutsoupias and Papadimitriou in 1999 [108], but the idea of measuring inefficiency of equilibrium is older (Dubey, 1986 [43]).

The PoA is a measure of inefficiency due to the removal of central control and the introduction of selfish behaviour. It is defined as the ratio of the highest Nash cost to optimal cost in a game where players aim to reduce costs. In this thesis the overall objective is to maximise throughput; thus the PoA is equivalently considered as the ratio of the optimal social welfare to the lowest welfare at Nash equilibrium. Since the work included in this thesis concentrates only on pure strategy Nash equilibria, only the pure PoA will be considered (in effect the reciprocal of the welfare as a cost function is used in this thesis). Thus for the prisoners dilemma the PoA is $2 = 4/2$. Of course, the Nash equilibrium may not always exist; in this thesis the convention will be used that in the absence of Nash equilibria in pure strategies, the PoA tends to infinity.

7.2.1 Game Theory in Healthcare

Most research where game theory is applied in healthcare has mainly concentrated on Emergency Departments (EDs) and how to deal with diversions of patients and ambulances. Hagtvedt *et al.*, 2009 [78] considered cooperative strategies for hospitals, in order to reduce occurrences when ambulances are turned away due to the ED being full. These strategies lie in between the extreme of individualism and a central planner approach. The game theory approach was used to show that without some form of cooperative scheme, the incentives to defect are strong and will often lead

to system-wide pre-emptive diversion. The hospitals operate under a Prisoners' Dilemma when the patient load is sufficient. They claimed that the penalty must be large enough to balance out the variance of the overflow before there is a an optimal threshold. Having examined these aspects of cooperative solutions to ambulance diversion, they believed the incentives require a centralised agent to route patients, at least when the patient load is high.

Deo and Gurvich, 2011 [41] proposed a queueing network model of two EDs to study the network effect of ambulance diversion. Each ED aims to minimise the expected waiting time of its patients (walk-ins and ambulances) and chooses its diversion threshold based on the number of patients at its location. They modelled the decentralised decision making in the network as a non-cooperative game. Analysis of the game reveals that, at equilibrium, EDs declare diversion status defensively to avoid getting arrivals from each other. This equilibrium undermines all potential pooling benefits of ambulance diversion, a phenomenon labelled as the depooling effect. These results provided one potential explanation for the evidence regarding defensive diversion and the impact of cancelling ambulance diversion in Massachusetts in January 2009. They proposed and analysed an alternative solution to the social planner's problem in which the diversion thresholds are set to be equal to the EDs' respective capacities. When there are available beds in one ED simultaneously with queued patients at the other, this policy routes all the "refutable" patients to the ED with available beds and thus recovers most of the pooling benefits. In addition to its being easier to implement than the true social optimum, it reduced the expected waiting times of both EDs.

Knight, 2012 [102] in his working paper considered a non-cooperative game, modelling a system of two hospitals and two interacting services: the Emergency Medical Vehicle and the Emergency Department. Using the PoA measure he showed that high levels of inefficiencies can be obtained due to the hospitals acting selfishly.

Some other work that has not concentrated on EDs, but has healthcare implications includes: Knight and Harper, 2012 [103] work, where results concerning the congestion related implications of decisions made by patients when choosing between healthcare facilities were presented. Using theoretical results from routing game theory the following conclusions regarding the PoA were proved analytically: the PoA increases with worth of service, up to a point; in a system with insufficient capacity the PoA is low; and choice causes the highest level of inefficiency when the capacity of the system matches the perceived worth of service.

Game theory may be useful in modelling patient and doctor arrivals by considering the conflicting interests of both parties. It is likely that patients arrive early to beat the system or arrive late knowing that they will have to wait anyway. Similarly, doctors may arrive late, assuming that the first patient will be late. There should be either some sort of mechanisms to enforce punctuality, or the

appointment systems should be designed to account for all parties behaviour (Van Ackere, 1990 [159]). One might expect that when clinics are run under more credible appointment systems, both patients and doctors will become more punctual.

Howard, 2002 [88] developed a model of the accept / reject decision for transplant organs and he showed how game theory might be used in making risk / benefit decisions in diagnostic radiology and other areas where risk / benefit needs to be considered. Howard believed that physicians would reject a low-quality organ if the patient is relatively healthy and can wait for a better quality organ.

Roth has made significant contributions to the fields of game theory and is known for his emphasis on applying his economic theory to solutions for "real-world" problems, including healthcare. Roth developed a very successful clearinghouse to facilitate the matching of doctors to residence programs (Roth, 1984 [142]). Today this clearinghouse is called the National Resident Matching Program. Roth along with Sonmez and Unver is a founder of the New England Program for Kidney Exchange that pairs compatible kidney donors and recipients. Roth in 2012 won the Nobel Memorial Prize in Economic Sciences jointly with Shapley "for the theory of stable allocations and the practice of market design".

Most of existing game-theoretic queueing models focused on a setting in which the firms' decision is either price / or capacity (Levhari and Luski, 1978 [111]; Cachon and Harker, 2002 [20]; Kalai *et al.*, 1992 [96]; Cachon and Zhang, 2007 [21]; Allon and Federgruen, 2007 [4]). In these models, the choice of price / or capacity determines the arrival rate for each firm. The work by: Tezcan, 2008 [156], Stolyar, 2005 [153], and Adan *et al.*, 1994 [1] is also not directly applied in healthcare settings; however the authors studied settings in which routing decisions are made upon customer / patient arrival, and once the customers are assigned to a queue they cannot be rerouted, therefore their models can be easily applied to different healthcare models.

7.3 Introduction

As mentioned before, this part of the thesis will concentrate on a non-cooperative game theoretical models of two CCUs. It is assumed that both CCUs act selfishly and the impact of lack of collaboration will be studied. If CCUs are overcrowded they can declare being in "transfer" status and patients are diverted to the other CCU if they have available beds to accept extra patients.

It is assumed that the players are the CCU managers, who assess crowding in terms of the total number of patients in the CCU and request transfer when this crowding measure exceeds a predetermined cut-off. The other CCU will accept the transfer request if their bed occupancy is below their predetermined cut-off. Otherwise, the transfer request is cancelled and depending on the

model, either each CCU is forced to accept its own patients or patients are refused admission to a CCU and are admitted to an ordinary ward within the hospital. Each player chooses a transfer cut-off with the objective of maintaining the utilisation rate as close to, but below, 80%. The 80% utilisation rate will be referred to as the target.

A Markov chain model of the two CCUs, the Nevill Hall (NH) and the Royal Gwent (RG) will be adapted to investigate the impact of patient's transfers. Each CCU will be admitting two arrival streams: their own patients and transfers from the other CCU.

7.4 Basic Methodology

To formally investigate the impact of decentralised decision making, the interaction between two CCUs is placed within a non-cooperative game framework. The interaction will be modelled through a two dimensional Markov chain. In this section, the basic Markov model (with no actual interaction) will be described, before moving on to the next sections that modify the Markov chain.

With no transfers, each CCU faces only one arrival stream: their own patients arriving according to a Poisson process with rate λ_1 at NH and λ_2 at RG. It is assumed that the length of stay (LoS) of a patient is Exponentially distributed with mean $\frac{1}{\mu_1}$ at NH and $\frac{1}{\mu_2}$ at RG. Assume the CCU at NH has m beds, and the CCU at RG has n beds available. The underlying basic Markov chain for a network comprising two CCUs is illustrated in Figure 7.2.

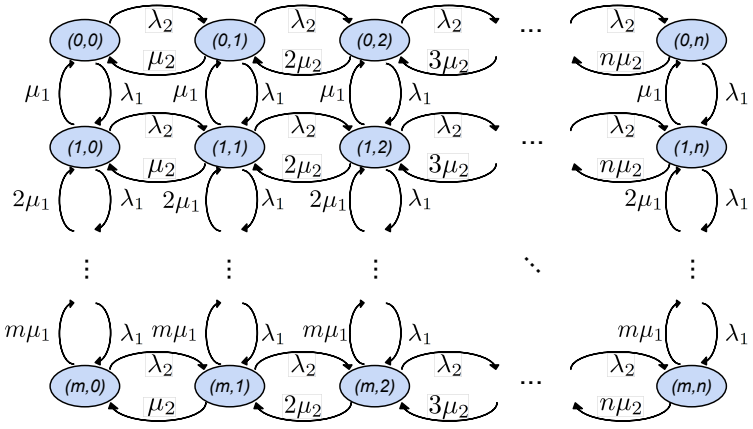


Figure 7.2: Basic Markov chain

Each state (u, v), where $u \in [0, \ldots, m]$ and $v \in [0, \ldots, n]$, denotes the situation where u beds are occupied in NH and v beds occupied in RG. In total there are $(m + 1) \times (n + 1)$ states and they are indexed lexicographically: $(0, 0), (0, 1), (0, 2)$, etc. In general, state (u, v) is the $s(u, v)th$ state

where $s(u,v) = u(n+1) + (v+1)$.

The stochastic transition rate matrix $Q = Q(m,n)$ of the continuous-time Markov chain ([152]) has the form:

$$Q = \begin{pmatrix} -\sum\limits_{j\neq i} q_{i,j} & q_{1,2} & q_{1,3} & \cdots & q_{1,(m+1)(n+1)} \\ q_{2,1} & -\sum\limits_{j\neq i} q_{i,j} & q_{2,3} & \cdots & q_{2,(m+1)(n+1)} \\ q_{3,1} & q_{3,2} & -\sum\limits_{j\neq i} q_{i,j} & \cdots & q_{3,(m+1)(n+1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{(m+1)(n+1),1} & q_{(m+1)(n+1),2} & q_{(m+1)(n+1),3} & \cdots & -\sum\limits_{j\neq i} q_{i,j} \end{pmatrix}$$

where $q_{i,j}$ is the rate at which a transition from state i to state j occurs. A matrix Q has non-negative off-diagonal elements; row sums are equal to zero and diagonal elements are equal to the negated sum of off-diagonal row elements. Each row has at most four non-zero elements: for the basic model, where transfers are not allowed, they are: $u\mu_1, v\mu_2, \lambda_1$ and λ_2. A diagrammatic representation is given in Figure 7.3.

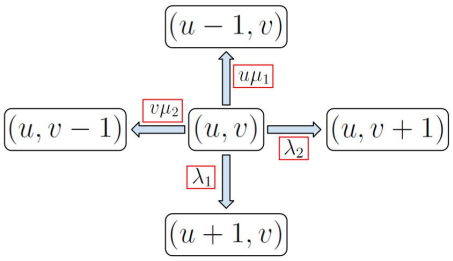


Figure 7.3: Possible transition rates between states for the basic Markov chain

The transition rates are given by:

$$q_{i,j} = \begin{cases} u\mu_1 & \text{if } (u_i,v_i) - (u_j,v_j) = (1,0), \\ v\mu_2 & \text{if } (u_i,v_i) - (u_j,v_j) = (0,1), \\ \lambda_1 & \text{if } (u_i,v_i) - (u_j,v_j) = (-1,0), \\ \lambda_2 & \text{if } (u_i,v_i) - (u_j,v_j) = (0,-1), \\ 0 & \text{otherwise.} \end{cases} \tag{7.1}$$

Having the transition rate matrix Q it is possible to obtain the stationary distribution of Markov chains from the system of linear equations:

$$\pi Q = 0, \qquad \pi \geq 0 \tag{7.2}$$

where

$$\pi = (\pi_1, \pi_2, \ldots, \pi_{(m+1)(n+1)})$$

and π_s is the probability of being in a state s. By transposing Equation 7.2:

$$Q^T \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_{(m+1)(n+1)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{7.3}$$

The standard direct approaches for solving systems of linear equations are based on the method of Gaussian elimination, which is implemented in *MATLAB* (and *Python* in Chapter 8 for reasons that will become clear) and pseudo code is available in Appendix E.

Having probabilities of being in a state s, it is possible to convert them to probabilities of being in a state $(f(s), g(s))$ using the following functions: $f(s) = \left\lfloor \frac{s-1}{n+1} \right\rfloor$ and $g(s) = (s-1) mod\ (n+1)$, where n is the RG bed capacity. Therefore, probabilities of a given bed occupancy for each hospital can be obtained using the following formulas:

$$P^{NH}(u) = \sum_{v=0}^{n} P(u, v) \qquad \text{for } u = 0, 1, \ldots, m$$

$$P^{RG}(v) = \sum_{u=0}^{m} P(u, v) \qquad \text{for } v = 0, 1, \ldots, n$$

7.5 Game Theoretic Model

The methodology described in this section will be used to solve game theoretic models in Section 7.7 and 7.8. The queueing network is embedded within a static non-cooperative game, where the two players are CCU managers, the first player being referred to as NH and the second as RG. Each player chooses a transfer cut-off with the objective of maintaining the utilisation rate as close to, but below 80%.
Let:

K_{NH}	the transfer cut-off at NH	$K_{NH} \in [0, \ldots, m]$
K_{RG}	the transfer cut-off at RG	$K_{RG} \in [0, \ldots, n]$

Let $P(h)$ be the probability of having h beds occupied; thus each CCU is faced with the following optimisation problem:

Minimize:

$$\left(\frac{\sum_{h=1}^{c} hP(h)}{c} - \text{target} \right)^2$$

Subject to:

$$0 \leq K \leq c$$

$$K \in \mathbb{Z}$$

$$\frac{\sum_{h=1}^{c} hP(h)}{c} \leq \text{target}$$

where: K is the cut-off, $c \in [m, n]$ is the capacity of the corresponding CCU.

The general arrival rates in each region separated by the cut-off points are given below for each CCU in Figure 7.4.

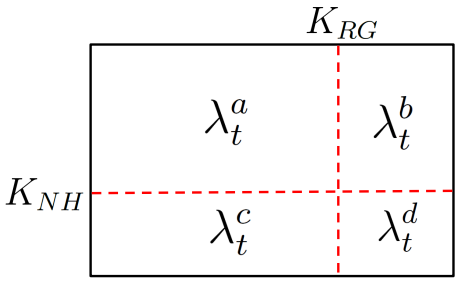


Figure 7.4: General arrival rates for each CCU at each region, where $t \in \{NH, RG\}$

For each possible choice of K_{NH}, RG picks K_{RG} for which the utilisation rate at RG is below the specified target. A set of best RG responses is created: $B_{RG}(K_{NH})$. In a similar way, for each possible choice of K_{RG}, NH picks K_{NH}, for which the utilisation rate at NH satisfy the utilisation target. A set of best NH responses $B_{NH}(K_{RG})$ to each RG cut-off is also created. Having a set of best responses for each CCU, a set of pure Nash equilibria can be obtained. Recall Section 7.2, the pair $(\widehat{K_{NH}}, \widehat{K_{RG}})$ is said to be the Nash equilibrium if and only if $B_{RG}(\widehat{K_{NH}}) = \widehat{K_{RG}}$ and $B_{NH}(\widehat{K_{RG}}) = \widehat{K_{NH}}$.

7.6 Price of Anarchy

Throughput of patients is a natural choice of utility given that most hospitals are financially rewarded per served patient ([134]). For each cut-off pair (K_{NH}, K_{RG}) the utilisation rate and

throughput can easily be obtained for each CCU, using the following formulas:

$$\text{utilisation} = \frac{\sum_{h=0}^{c} hP(h)}{c}$$

$$\text{throughput} = \mu \sum_{h=0}^{c} hP(h)$$

where c is the bed capacity and h is the number of beds occupied in the corresponding CCU. The overall throughput of both CCUs is also obtained. Next, the maximum overall throughput is found, which will be referred to as the optimal throughput. Also, the throughput at Nash equilibrium is obtained. If there is more than one Nash equilibrium, the throughput for each Nash equilibria are compared and, by definition (Section 7.2), the lowest is picked and will be referred to as the Nash throughput.

The aim of the work presented is to measure the inefficiency created by the competitive interaction between CCUs. The approach is based on the Price of Anarchy (PoA), which is the ratio of the social optimum welfare to the welfare of the worst Nash equilibrium. That is, the ratio of the largest social welfare, T^* to the smallest social welfare, $\widehat{T}$, achieved at any Nash equilibrium. The social optimum in this case is the optimal throughput, hence:

$$\text{PoA} = \frac{T^*}{\widehat{T}} = \frac{\text{optimal throughput}}{\text{Nash throughput}}$$

Since the optimal throughput will always be greater or equal to the Nash throughput, the PoA is always greater or equal to 1; it tends to infinity in two occasions:

- if the Nash equilibrium does not exist;
- if throughput at the Nash equilibrium is equal zero.

All of the above is implemented in *MATLAB* (and *Python* in the next chapter for reasons that will become clear) due to numerical imprecisions occurring in VBA.

The queueing model described in Section 7.4 was basic; no interaction occurred between the CCUs. In Sections 7.7 and 7.8 this situation is modelled as a game by having transition rates dependent on the strategies of each CCU.

7.7 Model 1

Recalling Figure 7.4, this model assumes that $\lambda_1^c = \lambda_1^d = 0$ and $\lambda_2^b = \lambda_2^d = 0$, which means that if the bed occupancy level at both Units exceeds a predetermined cut-off, then the admission to CCU

is cancelled and patients will be admitted to a Ward within the hospital. If each CCU chooses their cut-off at zero, patients are not admitted at all, and, consequently both Units are closed. It is also assumed by the motion of transfers that transferred patients will be treated under the length of stay profile of the CCU they are admitted to. The Markov chain used to model this game is shown in Figure 7.5.

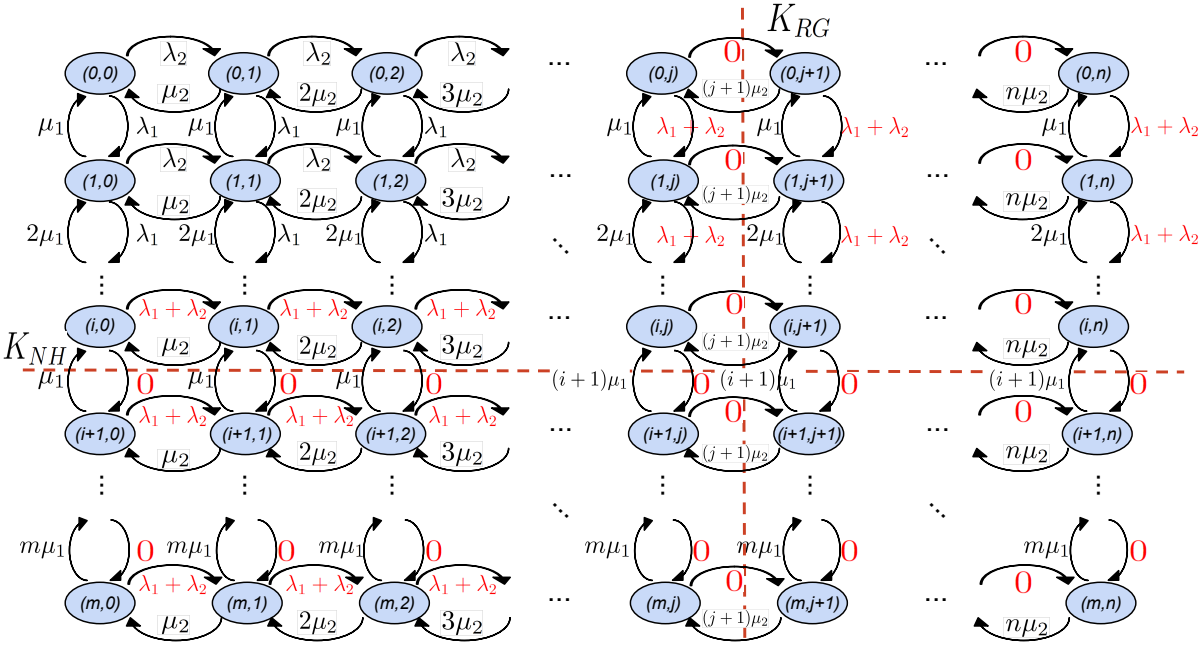


Figure 7.5: Markov chain for Model 1

Formally, the arrival rates ($\widetilde{\lambda}_1$ at NH and $\widetilde{\lambda}_2$ at RG) at each CCU are given by:

$$\widetilde{\lambda}_1 = \begin{cases} \lambda_1, & \text{if } i < K_{NH} \text{ and } j < K_{RG} \\ \lambda_1 + \lambda_2, & \text{if } i < K_{NH} \text{ and } j \geq K_{RG} \\ 0, & \text{if } i \geq K_{NH} \end{cases} \qquad \widetilde{\lambda}_2 = \begin{cases} \lambda_2, & \text{if } i < K_{NH} \text{ and } j < K_{RG} \\ \lambda_1 + \lambda_2, & \text{if } i \geq K_{NH} \text{ and } j < K_{RG} \\ 0, & \text{if } j \geq K_{RG} \end{cases}$$

Therefore, the matrix Q can be obtained from the following transition rates $q_{i,j}$:

$$q_{i,j} = \begin{cases} u_i\mu_1 & \text{if } (u_i, v_i) - (u_j, v_j) = (1, 0), \\ v_i\mu_2 & \text{if } (u_i, v_i) - (u_j, v_j) = (0, 1), \\ \lambda_1 & \text{if } (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } v_i < K_{RG}, \\ \lambda_2 & \text{if } (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } u_i < K_{NH}, \\ \lambda_1 + \lambda_2 & \text{if } \begin{cases} (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } u_i < K_{NH} \text{ and } v_i \geq K_{RG} \text{ or} \\ (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } u_i \geq K_{NH} \text{ and } v_i < K_{RG}, \end{cases} \\ 0 & \text{otherwise.} \end{cases} \tag{7.4}$$

The following set of parameters, obtained from the data, will be used:

Table 7.2: Parameter values used in the model

Parameter	**Parameter description**	**Parameter value**
m	the bed capacity at NH	8
n	the bed capacity at RG	16
λ_1	the arrival rate at NH (per day)	1.498
λ_2	the arrival rate at RG (per day)	2.245
μ_1	the service rate at NH (days)	0.262
μ_2	the service rate at RG (days)	0.198
target	bed utilisation target	80%

For a given set of parameters, the best responses to a given cut-off for each CCU are obtained, and are illustrated in Figure 7.6.

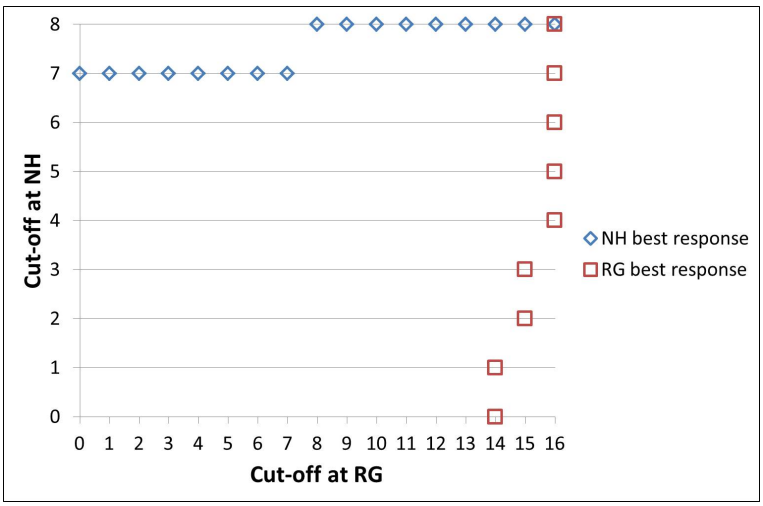
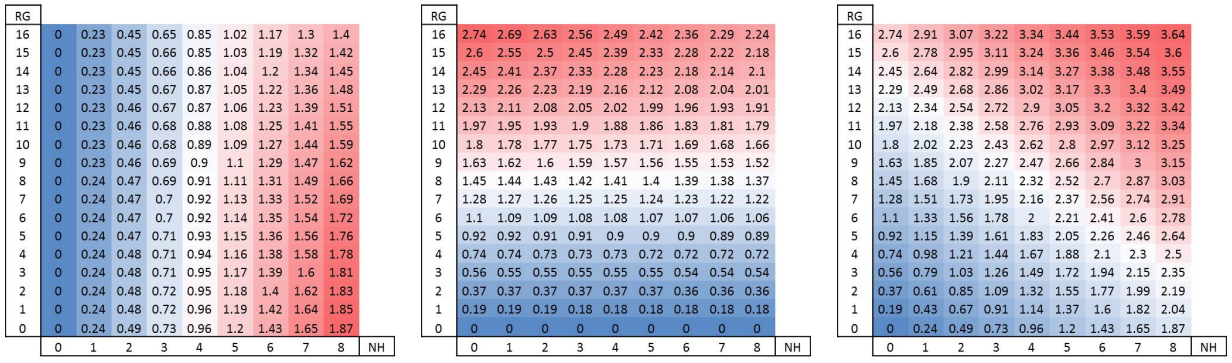


Figure 7.6: Best responses for each hospital

For example, if the RG picks $K_{RG} = 6$, NH should choose $K_{NH} = 7$. Similarly, if the NH picks $K_{NH} = 3$, RG should choose $K_{RG} = 15$. The two best responses overlap at $(8, 16)$, meaning that the Nash equilibrium is at $(8, 16)$. Having the Nash equilibrium at both CCUs full bed capacity means that if neither CCU turn away their patients, both Units will be running at a utilisation rate of less than 80%. If the utilisation rate target is lowered, CCUs turn away patients earlier; for example, if the target was changed to 70% or 60%, the Nash equilibrium would be at $(8, 15)$ or $(6, 11)$ respectively.

The throughput for each pair of cut-off points at each CCU (Figures 7.7a and 7.7b) and the overall throughput (Figure 7.7c) is evaluated.



RG									
16	0	0.23	0.45	0.65	0.85	1.02	1.17	1.3	1.4
15	0	0.23	0.45	0.66	0.85	1.03	1.19	1.32	1.42
14	0	0.23	0.45	0.66	0.86	1.04	1.2	1.34	1.45
13	0	0.23	0.45	0.67	0.87	1.05	1.22	1.36	1.48
12	0	0.23	0.46	0.67	0.87	1.06	1.23	1.39	1.51
11	0	0.23	0.46	0.68	0.88	1.08	1.25	1.41	1.55
10	0	0.23	0.46	0.68	0.89	1.09	1.27	1.44	1.59
9	0	0.23	0.46	0.69	0.9	1.1	1.29	1.47	1.62
8	0	0.24	0.47	0.69	0.91	1.11	1.31	1.49	1.66
7	0	0.24	0.47	0.7	0.92	1.13	1.33	1.52	1.69
6	0	0.24	0.47	0.7	0.92	1.14	1.35	1.54	1.72
5	0	0.24	0.47	0.71	0.93	1.15	1.36	1.56	1.76
4	0	0.24	0.48	0.71	0.94	1.16	1.38	1.58	1.78
3	0	0.24	0.48	0.71	0.95	1.17	1.39	1.6	1.81
2	0	0.24	0.48	0.72	0.95	1.18	1.4	1.62	1.83
1	0	0.24	0.48	0.72	0.96	1.19	1.42	1.64	1.85
0	0	0.24	0.49	0.73	0.96	1.2	1.43	1.65	1.87
	0	1	2	3	4	5	6	7	8	NH

(a) NH

RG									
16	2.74	2.69	2.63	2.56	2.49	2.42	2.36	2.29	2.24
15	2.6	2.55	2.5	2.45	2.39	2.33	2.28	2.22	2.18
14	2.45	2.41	2.37	2.33	2.28	2.23	2.18	2.14	2.1
13	2.29	2.26	2.23	2.19	2.16	2.12	2.08	2.04	2.01
12	2.13	2.11	2.08	2.05	2.02	1.99	1.96	1.93	1.91
11	1.97	1.95	1.93	1.9	1.88	1.86	1.83	1.81	1.79
10	1.8	1.78	1.77	1.75	1.73	1.71	1.69	1.68	1.66
9	1.63	1.62	1.6	1.59	1.57	1.56	1.55	1.53	1.52
8	1.45	1.44	1.43	1.42	1.41	1.4	1.39	1.38	1.37
7	1.28	1.27	1.26	1.25	1.25	1.24	1.23	1.22	1.22
6	1.1	1.09	1.09	1.08	1.08	1.07	1.07	1.06	1.06
5	0.92	0.92	0.91	0.91	0.9	0.9	0.9	0.89	0.89
4	0.74	0.74	0.73	0.73	0.73	0.72	0.72	0.72	0.72
3	0.56	0.55	0.55	0.55	0.55	0.55	0.54	0.54	0.54
2	0.37	0.37	0.37	0.37	0.37	0.37	0.36	0.36	0.36
1	0.19	0.19	0.19	0.18	0.18	0.18	0.18	0.18	0.18
0	0	0	0	0	0	0	0	0	0
	0	1	2	3	4	5	6	7	8	NH

(b) RG

RG									
16	2.74	2.91	3.07	3.22	3.34	3.44	3.53	3.59	3.64
15	2.6	2.78	2.95	3.11	3.24	3.36	3.46	3.54	3.6
14	2.45	2.64	2.82	2.99	3.14	3.27	3.38	3.48	3.55
13	2.29	2.49	2.68	2.86	3.02	3.17	3.3	3.4	3.49
12	2.13	2.34	2.54	2.72	2.9	3.05	3.2	3.32	3.42
11	1.97	2.18	2.38	2.58	2.76	2.93	3.09	3.22	3.34
10	1.8	2.02	2.23	2.43	2.62	2.8	2.97	3.12	3.25
9	1.63	1.85	2.07	2.27	2.47	2.66	2.84	3	3.15
8	1.45	1.68	1.9	2.11	2.32	2.52	2.7	2.87	3.03
7	1.28	1.51	1.73	1.95	2.16	2.37	2.56	2.74	2.91
6	1.1	1.33	1.56	1.78	2	2.21	2.41	2.6	2.78
5	0.92	1.15	1.39	1.61	1.83	2.05	2.26	2.46	2.64
4	0.74	0.98	1.21	1.44	1.67	1.88	2.1	2.3	2.5
3	0.56	0.79	1.03	1.26	1.49	1.72	1.94	2.15	2.35
2	0.37	0.61	0.85	1.09	1.32	1.55	1.77	1.99	2.19
1	0.19	0.43	0.67	0.91	1.14	1.37	1.6	1.82	2.04
0	0	0.24	0.49	0.73	0.96	1.2	1.43	1.65	1.87
	0	1	2	3	4	5	6	7	8	NH

(c) Overall

Figure 7.7: Throughput at each hospital for a given pair of cut-off points

The lowest throughput (blue) at individual CCUs is observed for their choice of low cut-off; the throughput gradually increases (red) as each CCU increases their cut off point. The highest throughput at an individual hospital occurs at maximum bed capacity and if the other hospital's cut-off is at zero. This is intuitive, since if one hospital is getting all admissions, their throughput will be highest. The highest overall throughput (Figure 7.7c) is achieved if neither of the CCUs turn away their patients i.e. at $(8, 16)$.

For this model there exists only one Nash equilibrium, at $(8, 16)$, at which Nash throughput is 3.64, the optimal throughput is also 3.64, therefore PoA = 1. Given the current rules (total rejections allowed and 80% target), selfish behaviour of hospitals does not effect efficiency of the system.

The model is investigated to check whether the 80% target can be decreased to still maintain a PoA of 1. As the target is increased from 10% to 100%, the PoA is non-increasing. The lowest target for which PoA=1 is 72%.

Furthermore, the PoA will be calculated for a few scenarios where the altered variables will be:

- target and the same percentage change in demand at both CCUs;
- demand at each CCU;
- bed capacity at each CCU.

7.7.1 'What if': Target and Percentage Demand Change

This Section will further investigate how the PoA changes for different targets along with different demand change at both hospitals. The PoA will be tested for:

- target $\in [0.1, 1]$ in steps of 0.05

- demand change, $x \in [-0.9, 2]$ in steps of 0.1
  resulting in: $\widetilde{\lambda}_1 = \lambda_1 \times (1 + x)$ and $\widetilde{\lambda}_2 = \lambda_2 \times (1 + x)$

Figure 7.8 presents how the PoA varies for different target values and demand rate changes.

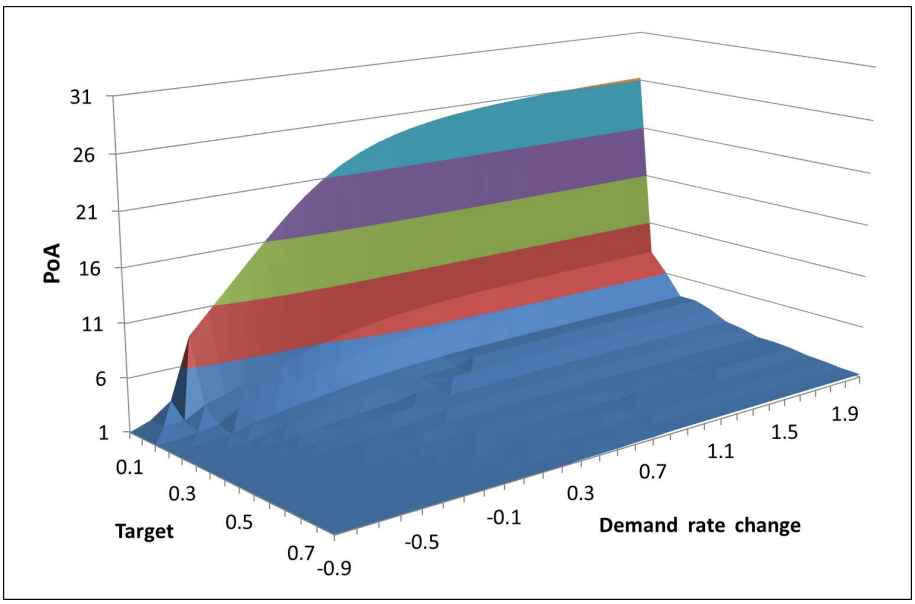


Figure 7.8: PoA for different target and demand rates

Tables 7.3 and 7.4 present the throughput at the Nash equilibrium and the optimal throughput correspondingly for a subsection of parameters.

Table 7.3: Nash throughput

		Demand rate change				
		0	**0.1**	**0.2**	**0.3**	**0.4**
Target	**0.55**	2.80	2.72	2.56	2.59	2.61
	0.6	3.09	2.87	2.93	2.97	3.01
	0.65	3.20	3.32	3.27	3.13	3.17
	0.7	3.60	3.58	3.41	3.49	3.54
	0.75	3.64	3.85	3.82	3.81	3.70

Table 7.4: Optimal throughput

		Demand rate change				
		0	**0.1**	**0.2**	**0.3**	**0.4**
Target	**0.55**	3.64	3.91	4.14	4.31	4.46
	0.6	3.64	3.91	4.14	4.31	4.46
	0.65	3.64	3.91	4.14	4.31	4.46
	0.7	3.64	3.91	4.14	4.31	4.46
	0.75	3.64	3.91	4.14	4.31	4.46

For any target value, as the demand increases, the PoA increases. For example, for a target of 0.65, when the total demand rate is 4.49 (20% increase) then the Nash equilibrium is at $(6, 11)$ resulting in a Nash throughput of 3.27 (Table 7.3), but the optimal throughput is 4.14 (Table 7.4) giving PoA=1.26. However, at a total demand rate of 4.86 (30% increase) the Nash equilibrium is at $(5, 11)$ at which the throughput is 3.13 (Table 7.3) and optimal is 4.31 (Table 7.4), giving

PoA=1.38. A higher demand causes CCUs to reject patients earlier, hence a higher rejection rate, therefore more inefficiency is observed.

The lower the target, the rate of change of the PoA is greater. As the target increases, the PoA levels get lower. A very high PoA is observed for a high demand and low target. Many patients arrive, but not many are admitted, hence a very high rejection rate, therefore a high inefficiency is observed.

In general, the PoA increases with demand and decreases with the target value. For 100% target the PoA is always 1, even for high demand increase. This is to be expected as setting a 100% target implies that both hospitals will try to maximise throughput.

7.7.2 'What if': Demand Change at Each CCU

In the previous scenario, the arrival rates were changed by the same percentage in both hospitals. Now, the PoA is calculated for different $\lambda_1 \in [0.1, 4.5]$ and $\lambda_2 \in [0.1, 7]$ in steps of 0.1, resulting in up to a triple increase of the original arrival rates. A target of 80% is used. Figure 7.9 demonstrates how the PoA varies.

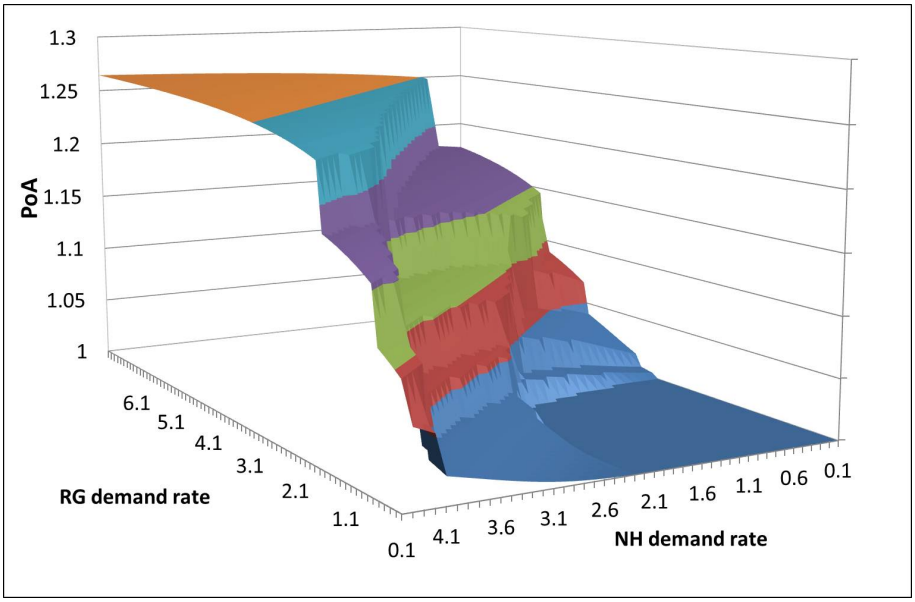


Figure 7.9: PoA for different demand rates at each hospital

Tables 7.5 and 7.6 present the throughput at the Nash equilibrium and the optimal throughput correspondingly for a subsection of parameters.

Table 7.5: Nash throughputs

NH demand rate	RG demand rate				
	2.8	**2.9**	**3**	**3.1**	**3.2**
1.3	3.84	3.89	3.95	4.00	3.95
1.4	3.90	3.95	4.00	3.96	4.00
1.5	3.95	4.00	4.05	4.00	4.036
1.6	4.00	4.05	4.10	4.04	4.07
1.7	4.06	4.10	4.04	4.07	3.95

Table 7.6: Optimal throughputs

NH demand rate	RG demand rate				
	2.8	**2.9**	**3**	**3.1**	**3.2**
1.3	3.90	3.96	4.02	4.09	4.13
1.4	3.96	4.02	4.08	4.13	4.18
1.5	4.02	4.08	4.14	4.19	4.23
1.6	4.08	4.14	4.19	4.24	4.28
1.7	4.14	4.19	4.24	4.28	4.33

The PoA increases with demand. For example, for the original NH arrival rate of 1.5 per day and arrival rate of 3 at RG, the throughput at Nash equilibrium at $(8, 15)$ is 4.05 (Table 7.5) and the optimal is 4.14 (Table 7.6), implying a PoA of 1.02. The throughput at individual CCUs at Nash equilibrium is 1.56 at NH and 2.49 at RG. If the arrival rate at RG is increased to 3.1, the throughput at Nash equilibrium $(8, 14)$ is 4.00 (Table 7.5) and the optimal is 4.19 (Table 7.6), implying a PoA of 1.05. The throughput at individual CCUs at Nash equilibrium is 1.52 at NH and 2.48 at RG. Therefore a slight increase in arrival rate:

- makes both CCUs transfer patients earlier;
- decreases throughout at Nash equilibrium at both CCUs;
- increases rejection rate;

For very high demand at both CCUs, rejections of patients are unavoidable. If $\lambda_1 = 4.5$ and $\lambda_2 = 7$, the PoA is highest (1.26). A PoA of 1.26 corresponds to 26% patients rejected, which is not an acceptable quantity.

7.7.3 'What if': Bed Capacity Change at Both CCUs

This scenario will investigate the effect on the PoA of changing the number of available beds at both CCU. The capacity at NH will vary from 1 to 16 and from 1 to 32 at RG. Figure 7.10 presents the PoA for the various capacities.

Tables 7.7 and 7.8 present the throughput at the Nash equilibrium and the optimal throughput correspondingly for a subsection of parameters.

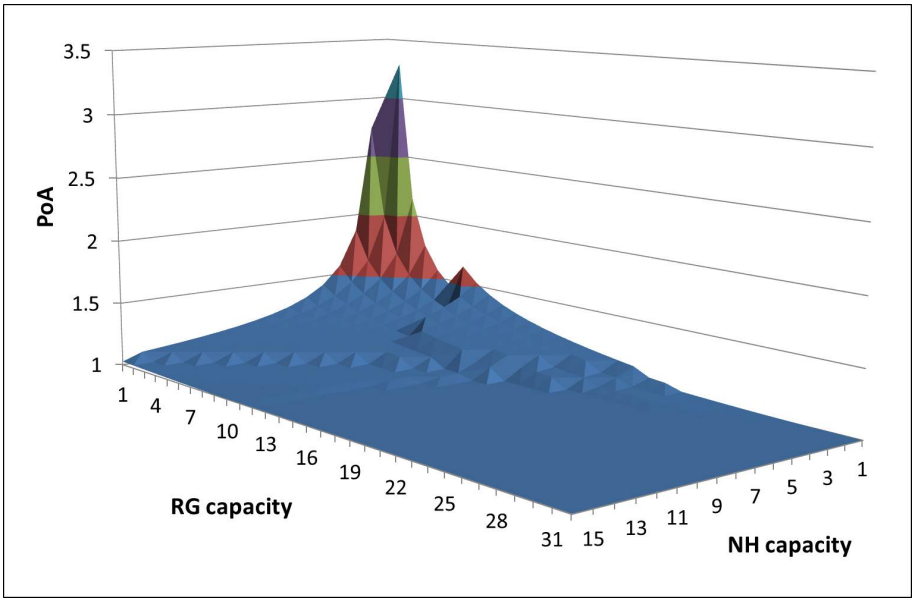


Figure 7.10: PoA for different bed capacity at each CCU

Table 7.7: Nash throughputs

		RG capacity				
		10	**11**	**12**	**13**	**14**
NH capacity	**6**	2.66	2.80	3.09	3.20	3.38
	7	2.84	3.12	3.22	3.32	3.48
	8	3.15	3.25	3.34	3.49	3.55
	9	3.27	3.36	3.44	3.57	3.61
	10	3.38	3.46	3.58	3.62	3.65

Table 7.8: Optimal throughputs

		RG capacity				
		10	**11**	**12**	**13**	**14**
NH capacity	**6**	2.97	3.09	3.20	3.30	3.38
	7	3.12	3.22	3.32	3.40	3.48
	8	3.25	3.34	3.42	3.49	3.55
	9	3.36	3.44	3.51	3.57	3.61
	10	3.46	3.52	3.58	3.62	3.65

In general, the PoA decreases with capacity increase. As an example, for a bed capacity at NH of 8 and 12 at RG, the throughput achieved at the Nash equilibrium of $(8, 11)$ is 3.34 (Table 7.7) and the optimum is 3.42 (Table 7.8), giving the PoA of 1.02. If the bed capacity is increased by 1 at RG, the Nash equilibrium is at $(8, 13)$, which means that neither of the CCUs would reject their patients. Having the Nash equilibrium at CCUs' full capacities means that the throughput at Nash is equal to the optimal throughput and PoA=1.

In a centralised system, for a NH bed capacity of 8, as currently, RG could reduce their bed capacity to 13, and similarly for a RG bed capacity of 16, NH could reduce their to 5, to maintain efficiency of the system with the PoA of 1. In a decentralised system that would not be the case.

There is a visible tipping point for a low NH capacity, between 1 and 5, and RG capacity between 7 and 8. Investigation of best responses showed that, for example, if NH had 2 beds and RG had 7 or 8 beds, the Nash equilibrium would be the same: $(1, 6)$; therefore the throughput at the Nash equilibrium remains the same; however, the optimum throughput increases for 8 beds at RG. As the optimum throughput increases, so does the PoA. To better understand the reason why RG does not increase their transfer cut-off when their bed capacity is increased, the utilisation rates are investigated. It appears that for bed capacities of 2 and 7 respectively in NH and RG, the utilisation rate at Nash equilibrium $(1, 6)$ is 79.84%. If an extra bed is added to RG, giving 8 in total, and RG started rejecting patients at $K_{RG} = 7$, the utilisation rate would be 81.13%, which does not satisfy the target constraint. For $K_{RG} = 6$, the utilisation rate is 69.86%, which is below the 80% target. This has also been the case for the remaining points in that region.

This shows that in certain situations the increase of capacity could increase inefficiency in a decentralised system. This is in essence an instance of Braess's Paradox (Braess, 1968 [17] (original version) and Braess *et al.*, 2005 [18] (translated to English version)). The result of Braess has been used before in healthcare setting by Gallivan and Utley, 2003 [55], where the authors identified potential operational problems that might occur once the patient choice is introduced.

In this model, there is the potential for both CCUs to reject patients at the same time, and so patients are lost to the entire system. The following model will investigate the effect of not allowing total rejections on the PoA.

7.8 Model 2

Recalling Figure 7.4, this model assumes that $\lambda_1^c = 0$ and $\lambda_2^b = 0$, which means that if bed occupancy levels at both Units exceed a pre-determined cut-off, then transfers are not allowed and each CCU has to accommodate their own patients. If each CCU choose a cut-off at zero, patients are not transferred at all, and, consequently, both Units admit their own patients. The Markov chain used to model this game is shown in Figure 7.11

The overall arrival rates ($\widetilde{\lambda}_1$ at NH and $\widetilde{\lambda}_2$ at RG) at each CCU are as follows:

$$\widetilde{\lambda}_1 = \begin{cases} \lambda_1, & \text{if } \begin{cases} i < K_{NH} \text{ and } j < K_{RG} \text{ or} \\ i \geq K_{NH} \text{ and } j \geq K_{RG} \end{cases} \\ \lambda_1 + \lambda_2, & \text{if } i < K_{NH} \text{ and } j \geq K_{RG} \\ 0, & \text{if } i \geq K_{NH} \text{ and } j < K_{RG} \end{cases} \quad \widetilde{\lambda}_2 = \begin{cases} \lambda_2, & \text{if } \begin{cases} i < K_{NH} \text{ and } j < K_{RG} \text{ or} \\ i \geq K_{NH} \text{ and } j \geq K_{RG} \end{cases} \\ \lambda_1 + \lambda_2, & \text{if } i \geq K_{NH} \text{ and } j < K_{RG} \\ 0, & \text{if } i < K_{NH} \text{ and } j \geq K_{RG} \end{cases}$$

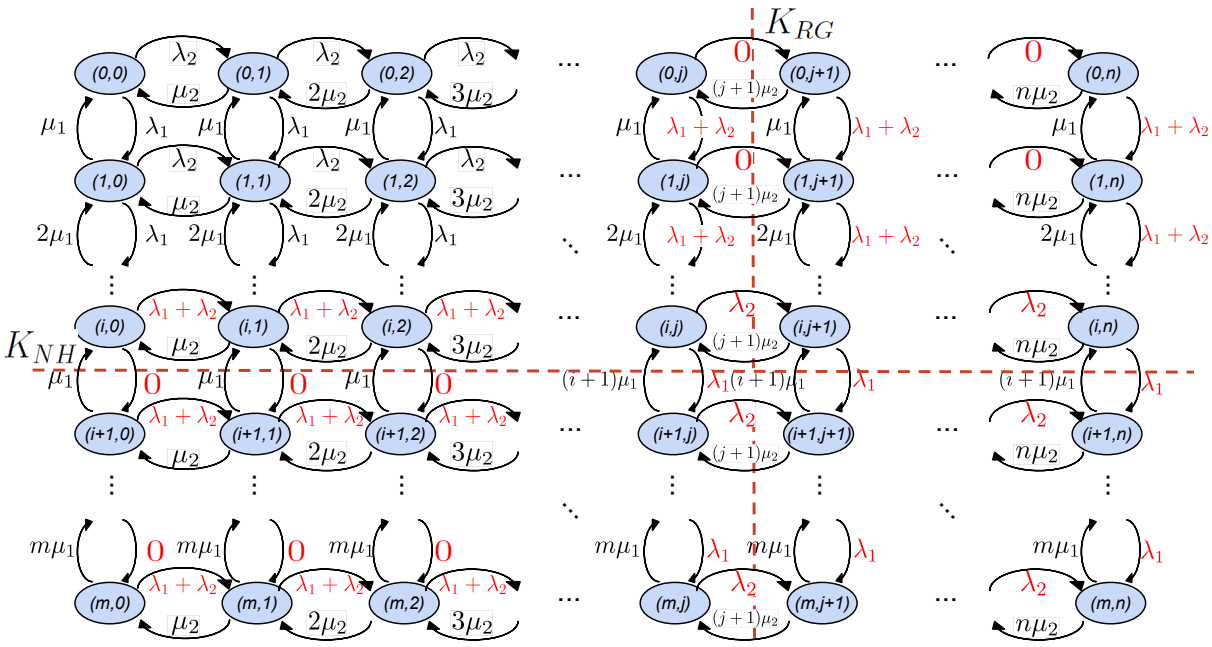


Figure 7.11: Markov chain for Model 2

Therefore, the transition matrix Q is obtained for the following transition rates $q_{i,j}$:

$$q_{i,j} = \begin{cases} u_i\mu_1 & \text{if } (u_i, v_i) - (u_j, v_j) = (1, 0), \\ v_i\mu_2 & \text{if } (u_i, v_i) - (u_j, v_j) = (0, 1), \\ \lambda_1 & \text{if } (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } \begin{cases} u_i < K_{NH} \text{ and } v_i < K_{RG} \text{ or} \\ u_i \geq K_{NH} \text{ and } v_i \geq K_{RG}, \end{cases} \\ \lambda_2 & \text{if } (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } \begin{cases} u_i < K_{NH} \text{ and } v_i < K_{RG} \text{ or} \\ u_i \geq K_{NH} \text{ and } v_i \geq K_{RG}, \end{cases} \\ \lambda_1 + \lambda_2 & \text{if } \begin{cases} (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } u_i < K_{NH} \text{ and } v_i \geq K_{RG} \text{ or} \\ (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } u_i \geq K_{NH} \text{ and } v_i < K_{RG}, \end{cases} \\ 0 & \text{otherwise.} \end{cases} \tag{7.5}$$

The same set of parameter values are used as in Model 1. The best responses for each CCU, given the chosen cut-off at other CCU, are presented in Figure 7.12.

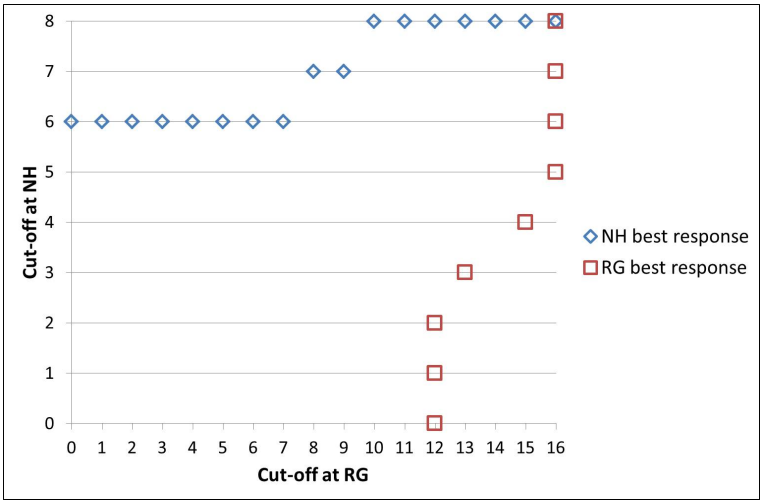


Figure 7.12: Best responses for each hospital for Model 2

The Nash equilibrium is at $(8, 16)$, as in Model 1, resulting in a PoA of 1. In comparison with Model 1, the individual best responses are decreased for low cut-off points at the competing CCU. For example, in Model 1, $B_{RG}(K_{NH} = 3) = 15$, in this model $B_{RG}(K_{NH} = 3) = 13$. This is due to the fact that cancellations are not allowed and the targets are reached faster.

The PoA is tested for different target values $\in [10\%, 100\%]$ in steps of 10% and $\in [60\%, 80\%]$ in steps of 1%. As the target increases, the PoA decreases. The minimum target for which PoA=1 is 72%. For a 70% target in Model 1 there was only one Nash equilibrium; in this model there are three pure Nash equilibria: $(6, 12)$, $(7, 13)$ and $(8, 14)$.

Furthermore, the throughputs for each pair of cut-off points, at each CCU (Figure 7.13a and 7.13b) and the overall throughput (Figure 7.13c) are evaluated.

RG	0	1	2	3	4	5	6	7	8	NH
16	0.40	0.48	0.62	0.77	0.92	1.07	1.21	1.32	1.40	
15	0.63	0.68	0.76	0.88	1.01	1.13	1.25	1.35	1.43	
14	0.80	0.82	0.88	0.97	1.08	1.19	1.30	1.39	1.46	
13	0.92	0.94	0.98	1.06	1.15	1.25	1.35	1.44	1.50	
12	1.03	1.04	1.07	1.13	1.21	1.31	1.40	1.48	1.54	
11	1.11	1.12	1.15	1.20	1.27	1.36	1.45	1.53	1.59	
10	1.18	1.19	1.21	1.25	1.32	1.41	1.50	1.58	1.63	
9	1.24	1.24	1.26	1.30	1.37	1.45	1.54	1.62	1.67	
8	1.28	1.28	1.30	1.34	1.40	1.49	1.58	1.66	1.72	
7	1.31	1.31	1.33	1.36	1.43	1.51	1.61	1.70	1.76	
6	1.32	1.33	1.34	1.38	1.44	1.53	1.63	1.73	1.79	
5	1.33	1.34	1.35	1.39	1.46	1.55	1.65	1.75	1.82	
4	1.34	1.34	1.35	1.39	1.46	1.55	1.66	1.76	1.84	
3	1.34	1.34	1.36	1.40	1.46	1.55	1.66	1.77	1.86	
2	1.34	1.34	1.36	1.40	1.46	1.56	1.66	1.78	1.87	
1	1.34	1.34	1.36	1.40	1.46	1.56	1.67	1.78	1.87	
0	1.34	1.34	1.36	1.40	1.46	1.56	1.67	1.78	1.87	

(a) NH

RG	0	1	2	3	4	5	6	7	8	NH
16	2.74	2.71	2.65	2.59	2.52	2.44	2.37	2.30	2.24	
15	2.69	2.67	2.63	2.57	2.49	2.42	2.34	2.27	2.20	
14	2.62	2.61	2.57	2.52	2.45	2.37	2.29	2.22	2.14	
13	2.55	2.54	2.51	2.46	2.39	2.32	2.24	2.15	2.07	
12	2.47	2.46	2.44	2.39	2.33	2.26	2.17	2.08	1.99	
11	2.39	2.38	2.36	2.33	2.27	2.19	2.10	2.00	1.89	
10	2.32	2.32	2.30	2.26	2.20	2.13	2.03	1.92	1.79	
9	2.26	2.26	2.24	2.21	2.15	2.07	1.96	1.84	1.69	
8	2.21	2.21	2.19	2.16	2.10	2.01	1.90	1.76	1.58	
7	2.18	2.17	2.16	2.13	2.06	1.97	1.85	1.69	1.48	
6	2.15	2.15	2.14	2.10	2.04	1.94	1.81	1.63	1.38	
5	2.14	2.14	2.13	2.09	2.02	1.92	1.78	1.58	1.29	
4	2.14	2.13	2.12	2.08	2.02	1.91	1.76	1.55	1.22	
3	2.13	2.13	2.12	2.08	2.01	1.91	1.75	1.53	1.17	
2	2.13	2.13	2.12	2.08	2.01	1.90	1.75	1.52	1.14	
1	2.13	2.13	2.12	2.08	2.01	1.90	1.75	1.52	1.12	
0	2.13	2.13	2.12	2.08	2.01	1.90	1.75	1.52	1.12	

(b) RG

RG	0	1	2	3	4	5	6	7	8	NH
16	3.14	3.19	3.27	3.36	3.44	3.51	3.57	3.62	3.64	
15	3.33	3.35	3.39	3.44	3.50	3.55	3.59	3.62	3.63	
14	3.42	3.43	3.46	3.49	3.53	3.57	3.59	3.61	3.61	
13	3.47	3.48	3.49	3.52	3.54	3.57	3.59	3.59	3.57	
12	3.49	3.50	3.51	3.52	3.54	3.56	3.57	3.57	3.53	
11	3.50	3.50	3.51	3.52	3.54	3.55	3.55	3.54	3.48	
10	3.50	3.50	3.51	3.52	3.53	3.53	3.53	3.50	3.42	
9	3.50	3.50	3.50	3.51	3.52	3.52	3.50	3.46	3.36	
8	3.49	3.49	3.49	3.50	3.50	3.50	3.48	3.43	3.30	
7	3.48	3.48	3.49	3.49	3.49	3.49	3.46	3.39	3.23	
6	3.48	3.48	3.48	3.48	3.48	3.48	3.44	3.36	3.17	
5	3.47	3.47	3.48	3.48	3.48	3.47	3.43	3.33	3.11	
4	3.47	3.47	3.47	3.48	3.48	3.46	3.42	3.32	3.06	
3	3.47	3.47	3.47	3.48	3.48	3.46	3.42	3.30	3.02	
2	3.47	3.47	3.47	3.48	3.48	3.46	3.41	3.30	3.00	
1	3.47	3.47	3.47	3.48	3.48	3.46	3.41	3.30	2.99	
0	3.47	3.47	3.47	3.48	3.48	3.46	3.41	3.30	2.99	

(c) Overall

Figure 7.13: Throughput for a given pair of cut-off points for Model 2

The throughput pattern is fairly different to Model 1; the lowest throughput at the first hospital is observed if the first CCU choose a low cut-off and the second choose a high, since the second

hospital will be able to accommodate most of the first hospital's demand. The throughput increases as both CCUs choose their cut-off to be in similar capacity proportion. As expected, the highest is reached if the first CCU choose not to reject at all and the second CCU picked a low cut-off. The overall throughput has a very different pattern to Model 1 as well, where the lowest was for a low pair of cut-off points. In Model 2 it is the lowest if one of the CCUs choose to reject patients at low bed occupancy and the other choose not to reject patients at all. The optimal throughput is observed at full bed capacity of both CCUs; therefore, if both CCUs cooperated by not rejecting patients, the system would be efficient.

Next, the PoA will be calculated for the same 'what if' scenarios as in Model 1 to investigate whether the constraint of not allowing total rejections has an impact on the PoA.

7.8.1 'What if': Target and Percentage Demand Change

This Section will investigate how the PoA changes for different targets along with different demand change at both hospitals. As in Model 1, the PoA will be tested for:

- target $\in [0.1, 1]$ in steps of 0.05
- demand change, $d \in [-0.9, 2]$ in steps of 0.1
  resulting in: $\widetilde{\lambda}_1 = \lambda_1 \times (1 + d)$ and $\widetilde{\lambda}_2 = \lambda_2 \times (1 + d)$

Figure 7.14 presents the PoA for different target values and demand rate changes.

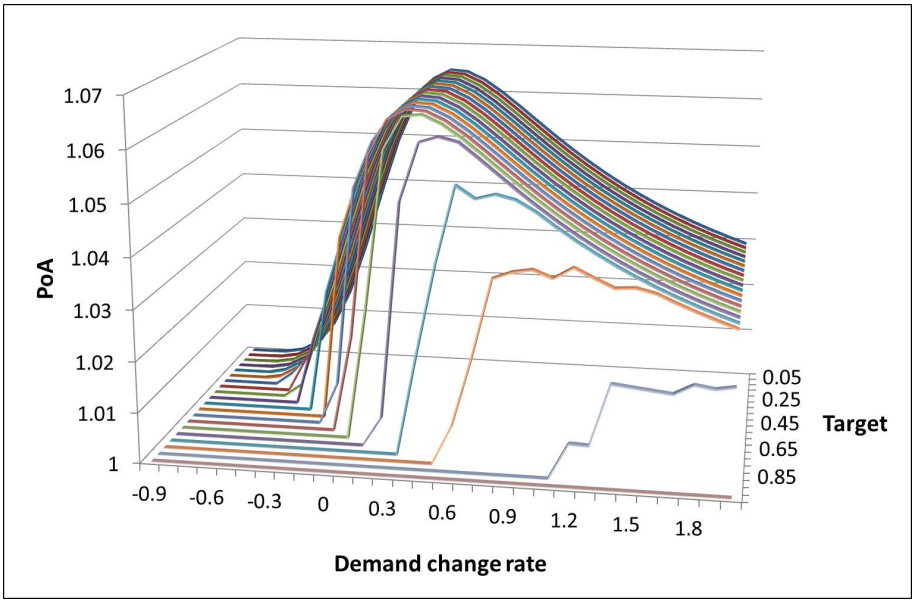


Figure 7.14: PoA for different target and demand rates for Model 2

Figure 7.14 illustrates the PoA in a 3D line graph; previously a surface 3D graph was used. Throughout the rest of this chapter, both types of graph will be used interchangeably depending on which one illustrates the situation more appropriately.

Figure 7.14 shows very interesting results; the PoA for a very small demand is low, as the system can easily deal with a small number of arrivals. As the demand change increases, the PoA increases; this is due to the fact that the optimal throughput increases much faster than the Nash throughput does. As the demand change increases even further, the PoA starts to decrease, because both CCUs have to deal with a very high demand and the difference between the optimal and Nash throughput gets smaller and smaller. The system is flooded and selfish behaviour does not have a negative effect; this is the expected behaviour as described in Knight and Harper, 2013 [103].

For example, for a target of 80% the PoA starts to rapidly increase for demand change higher than 0.1, and starts to decrease for demand change of 0.6; this region will be investigated closely. Table 7.9 presents results for 80% target and demand change from 0.1 to 0.6.

Table 7.9: Model 2 results for target of 80%

Demand change	**Nash equilibrium**	**Nash overall throughput**	**Nash NH throughput**	**Nash RG throughput**	**Optimal throughput**	**PoA**
0.1	(8,16)	3.91	1.51	2.41	3.91	1
0.2	(8,15)	4.11	1.63	2.48	4.14	1.06
0.3	(5,10)	4.11	1.63	2.48	4.31	1.07
0.4	(4,0)	4.20	1.65	2.55	4.46	1.06
0.5	(3,0)	4.30	1.66	2.64	4.57	1.06
0.6	(0,0)	4.39	1.68	2.71	4.65	1.06

Clearly, as the demand change increases, the Nash equilibria pairs decrease, because both CCUs are trying to transfer their patients earlier; if one CCU transfers early, the other will try as well, and as a result the Nash equilibrium for 0.6 demand increase is at $(0, 0)$, meaning that each CCU takes care of their own patients. As the demand increases even further the Nash equilibria remain at $(0, 0)$ and the PoA decreases; hence the increase in demand is not a problem, as the problem is in cooperation, or lack of it. Each CCU tries to be 'smarter' causing inefficiency in the whole system.

Assuming 100% target was acceptable, the arrival demand could be tripled to maintain PoA=1.

7.8.2 'What if': Demand Change at Each CCU

The model is now tested for a change in the arrival rates: $\lambda_1 \in [0.1, 4.5]$ and $\lambda_2 \in [0.1, 7]$ in steps of 0.1. The target of 80% is used. Figure 7.15 demonstrates how the PoA fluctuates.

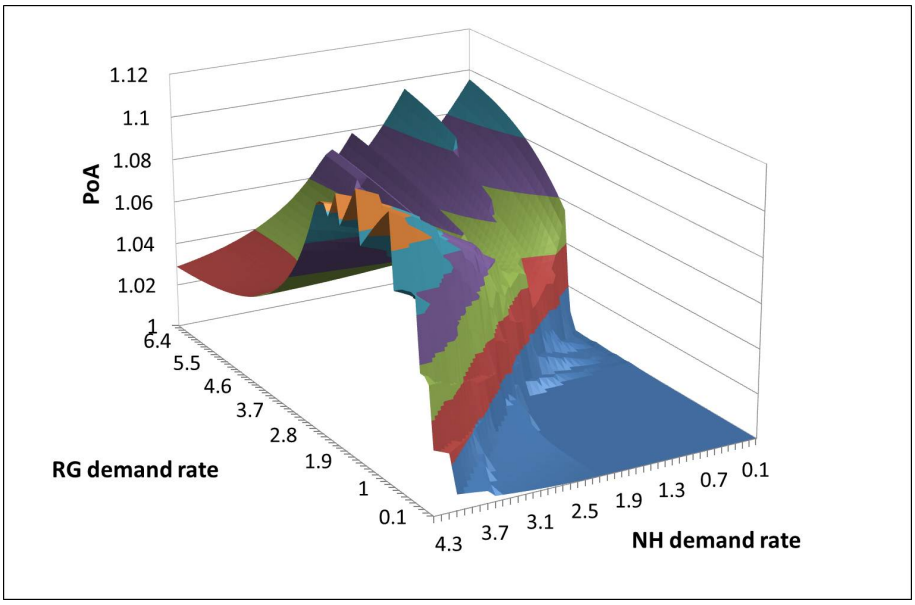
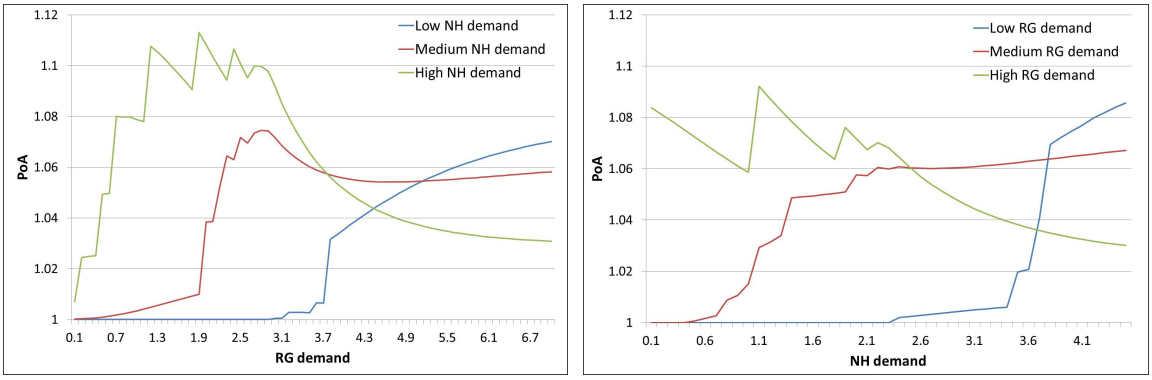


Figure 7.15: PoA for different demand rates at each hospital for Model 2

The PoA increases with demand; it remains at a value of 1 for all values in the region of $0.1 \leq \lambda_1 \leq 2$ and $0.1 \leq \lambda_2 \leq 2.4$. Due to the complexity of the graph, the PoA for each CCU, dependent on demand at the other CCU will be investigated separately (Figures 7.16a and 7.16b).



(a) NH (b) RG

Figure 7.16: Price of anarchy for each hospital as a function of other CCU demand

Both Figures (7.16a and 7.16b) show typical patterns for low, medium and high demand at individual CCUs. At the smaller CCU (NH), for low demand, the PoA stays low for low RG demand. At RG, the PoA for low demand also stays low for low NH demand; it also suddenly increases,

however, at a further point than NH. The bigger CCU can cope for longer with extra demand. The PoA for medium demand at NH, which is very close to the true demand, increases quite fast as RG demand increases, and then starts to decrease for large RG demand. Again, this is due to the fact that CCUs are less likely to make a worsening decision if they have to deal with large demand. At RG the medium demand, which is actually higher than the true demand, increases, but more steadily than at NH. The smaller CCU can handle their own high demand, but as soon as RG increases their demand the NH PoA rapidly increases to very high level, around 1.11, implying 11% of patients are rejected. For high RG demand (much higher than the true RG demand) PoA is high even for low NH demand, but after a certain point starts to decrease, for the same reason as in NH.

There are visible spikes on high demand in each CCU. As an illustration, an arrival rate of 4 at NH is taken and arrival rates at RG between 1.8 and 2.4 are investigated closely; the PoA will follow a similar pattern as shown in green in Figure 7.16a. Table 7.10 presents obtained results.

Table 7.10: Model 2 results for NH arrival rate of 4

RG arrival rate	**Nash equilibrium**	**Nash overall throughput**	**Nash NH throughput**	**Nash RG throughput**	**Optimal throughput**	**PoA**
1.8	$(0, 12)$	4.26	1.77	2.48	4.63	1.09
1.9	$(0, 12)$	4.29	1.79	2.50	4.65	1.08
2	$(0, 11)$	4.23	1.82	2.41	4.67	1.11
2.1	$(0, 11)$	4.26	1.83	2.43	4.69	1.10
2.2	$(0, 11)$	4.30	1.83	2.46	4.71	1.10
2.3	$(0, 11)$	4.33	1.84	2.49	4.73	1.09
2.4	$(0, 10)$	4.30	1.87	2.44	4.74	1.10

Since NH has a very high arrival rate, NH declares his cut-off to be at zero. So NH does not have to admit RG patients, and still RG is admitting a proportion of NH patients, therefore RG has to lower their cut-off to satisfy the 80% target constraint. The spikes are at the points where RG decreases their cut-off, initially from $K_{RG} = 12$ to $K_{RG} = 11$ and then to $K_{RG} = 10$. As an effect, Nash throughput at RG decreases, decreasing overall Nash throughput and therefore increasing the PoA.

In general, Figure 7.15 shows interesting and typical behaviour: the PoA is low for low demand at both CCU, increases for medium demand and starts to decrease for high demand at both CCUs.

7.8.3 'What if': Bed Capacity Change at Both CCUs

The final scenario will investigate the effect on the PoA by changing the number of available beds at both CCU. The bed capacity at NH will vary from 0 to 16 and at RG from 0 to 32. Figure 7.17

illustrates the PoA as a function of changed bed capacity at each CCU.

Figure 7.17 does not clearly show what the PoA for very low NH and very low RG occupancy is, hence the 3D line graph (Figure 7.18) is also shown.

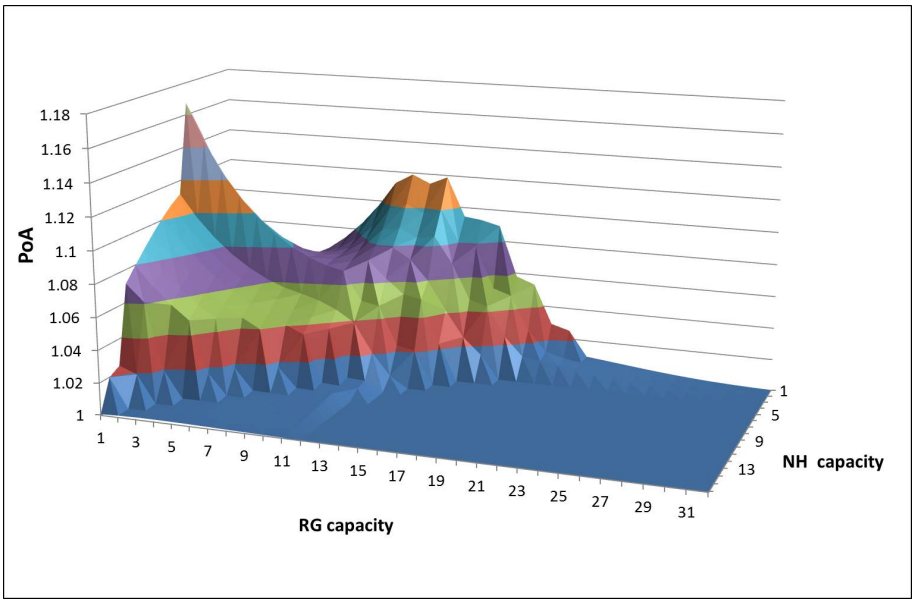


Figure 7.17: PoA for different bed capacity at each CCU for Model 2

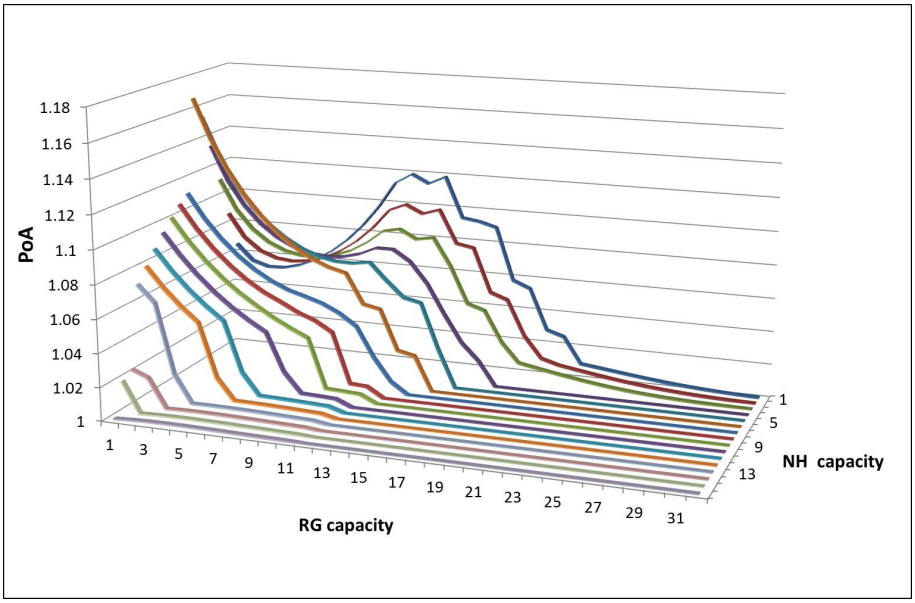


Figure 7.18: PoA for different bed capacity at each CCU for Model 2

The high PoA at both CCUs low bed capacities is expected, since patients are being lost due to an

insufficient number of beds for the current demand. In order to investigate the occurrence of the two tipping points, one at low RG capacity and NH capacity around 6 and the second at low NH capacity and RG capacity of around 13, attention will be given to Nash equilibria at both regions. The reason for both tipping points is in fact the same, and, as an example, for a bed capacity of 2 at NH closer inspection will be taken of the Nash equilibria as capacity in RG increases. Table 7.11 presents the results for RG capacity of between 10 and 16.

Table 7.11: Model 2 results for NH bed capacity of 2

RG bed capacity	**Nash equilibrium**	**Nash overall throughput**	**Nash NH throughput**	**Nash RG throughput**	**Optimal throughput**	**PoA**
10	(0,0)	2.05	0.44	1.62	2.23	1.09
11	(0,0)	2.17	0.44	1.74	2.38	1.10
12	(0,6)	2.30	0.44	1.87	2.54	1.10
13	(0,8)	2.44	0.42	2.02	2.68	1.09
14	(0,9)	2.56	0.42	2.14	2.82	1.10
15	(1,11)	2.73	0.42	2.31	2.95	1.08
16	(1,12)	2.85	0.41	2.44	3.07	1.08

At a low RG capacity the Nash equilibrium is $(0, 0)$, and as RG is receiving extra beds the equilibrium stays at the same level till RG has 11 beds. This is due to the fact that NH has only 2 beds available, so NH tries to transfer some patients to RG; however, since RG does not have a sufficient capacity to accommodate their own patients and a proportion of NH patients, RG chooses a cut-off at 0, so they only have to admit their own patients. As the RG capacity increases to 12, the cut-off chosen increases to 6, hence some NH patients are admitted to RG. As the Nash equilibria are increased, the overall throughput increases, decreasing the PoA.

7.9 Conclusions

A network consisting of two CCUs has been considered. A game theory approach has been used to quantify the effect of a decentralised system, where selfish behaviour of its agents caused inefficiency. To measure the ineffectiveness the price of anarchy has been used. It has been shown that in a healthcare system with an inadequate bed capacity providing the service, a high PoA is to be expected. In a system with the possibility of interaction, there is potential for anarchy and cooperation should not be neglected.

In general, Model 1 shows that by rejecting patients the system's efficiency is worsened. The PoA

for the corresponding 'what if' scenarios for Model 1 are much higher compared with Model 2. Also, Model 1 did not show the typical behaviour of PoA: it increases to a point, and then starts decreasing, which is due to the fact that at that point CCUs had too many strategies to pick from and the probability of picking a 'bad' tactic was high. After the tipping point CCUs either had enough capacity, so were not competing, or the demand was high, meaning that the system was 'bad', so the Nash and optimal throughput were not far apart.

The next Chapter will investigate more complex models, increasing the CCUs strategy choices even further.

Chapter 8

A Further Game Theoretical Consideration of Critical Care Unit Interaction

8.1 Introduction

In this chapter more detailed game theoretical models, where each Critical Care Unit (CCU) will have a much larger strategy space, will be proposed. As in Chapter 7, it is assumed that both CCUs act selfishly and inefficiency incurred as a result will be measured. In Model 1 (Section 7.7) and Model 2 (Section 7.8) it was assumed that, when the Units were overcrowded, patients were transferred to the other CCU. In this chapter, if CCUs are running on relatively high bed occupancy levels they declare being in "reduction" status first to later declare being in "transfer" status. The reduction status means that only a proportion of patients will be allowed to be admitted; the percentage of patients that were refused admission to the CCU will be admitted to an ordinary Ward. The transfer status means that patients will be diverted to the other CCU, assuming there are available beds to accept extra patients. As in Chapter 7, it is again assumed that patients obtain length of stay characteristics of the CCU they are treated in.

It is assumed that the players are the CCU managers, who assess crowding in terms of the total number of beds occupied in the CCU and declare reduction status when the number of patients exceeds the first predetermined cut-off level. If the number of occupied beds exceeds the second predetermined cut-off point then the players request transfer. The other CCU may accept the transfer request if their bed occupancy is below their second predetermined cut-off. Otherwise, the transfer request is cancelled, and depending on the model, either each CCU is forced to accept its own patients or patients are refused admission to the CCU. Each player chooses a reduction and transfer cut-off to keep the utilisation rate as close to, but below, 80%.

A more detailed Markov chain model of the two CCUs, the Nevill Hall (NH) and the Royal Gwent (RG), each admitting their own patients and transfers will be adapted to investigate the impact of patient's reductions and transfers.

8.2 Game Theoretic Model

The interaction will be modelled through a two dimensional Markov chain using the same methodology as described in Section 7.4 for Model 1 and Model 2. Assumptions regarding Poisson arrivals and Exponential service times are still valid. The bed capacity at each CCU remains the same, i.e. m in NH and n in RG; and the two players, as before, will be referred to as NH and RG. Each player chooses a reduction and transfer cut-off with the same objective as previously, to maintain the utilisation rate as close to, but below, 80%.

Let:

K^1_{NH}	the reduction cut-off at NH	$K^1_{NH} \in [0, \ldots, m]$
K^2_{NH}	the transfer cut-off at NH	$K^2_{NH} \in [K^1_{NH}, \ldots, m]$
K^1_{RG}	the reduction cut-off at RG	$K^1_{RG} \in [0, \ldots, n]$
K^2_{RG}	the transfer cut-off at RG	$K^2_{RG} \in [K^1_{RG}, \ldots, n]$

If for $j \in \{NH, RG\}$, $K^2_j = K^1_j$, then the problem reduces to Model 1 or Model 2 depending on whether the total rejections are allowed or not, and the reduction stage does not exist. Also, if $K^1_j = c$, where c is the bed capacity in the corresponding CCU, then the model reduces to the very basic model without any cut-off points, transfers are not allowed and each CCU admits only their own patients, as described in Section 7.4.

Let $P(h)$ be the probability of having h beds occupied; thus, each CCU is faced with the following optimisation problem:

Minimize:

$$\left(\frac{\sum_{h=1}^{c} hP(h)}{c} - \text{target} \right)^2$$

Subject to:

$$0 \leq K^1_j \leq K^2_j \leq c$$

$$K^1_j, K^2_j \in \mathbb{Z}$$

$$\frac{\sum_{h=1}^{c} hP(h)}{c} \leq \text{target}$$

where: K^1_j, K^2_j are the cut-off points, $j \in \{NH, RG\}$; c is the system capacity of the corresponding CCU.

For each possible choice of K^1_{NH}, K^2_{NH}, RG picks K^1_{RG}, K^2_{RG} for which the utilisation rate at RG is below the specified target. A set of pairs of best RG responses is created: $(B^1_{RG}(K^1_{NH}, K^2_{NH}), B^2_{RG}(K^1_{NH}, K^2_{NH}))$. In a similar way, for each possible choice of K^1_{RG}, K^2_{RG}, NH picks K^1_{NH}, K^2_{NH}, for which the utilisation rate at NH satisfy the utilisation target. A set of

best NH responses $(B^1_{NH}(K^1_{RG}, K^2_{RG}), B^2_{NH}(K^1_{RG}, K^2_{RG}))$ to each RG cut-off pair is also created. Having a set of pairs of best responses for each CCU, a set of pure Nash equilibria can be obtained. The quadruplet $\left(\widehat{K^1_{NH}}, \widehat{K^2_{NH}}, \widehat{K^1_{RG}}, \widehat{K^2_{RG}}\right)$ is said to be the Nash equilibrium if and only if:

- $B^1_{RG}(\widehat{K^1_{NH}}, \widehat{K^2_{NH}}) = \widehat{K^1_{RG}}$;
- $B^2_{RG}(\widehat{K^1_{NH}}, \widehat{K^2_{NH}}) = \widehat{K^2_{RG}}$;
- $B^1_{NH}(\widehat{K^1_{RG}}, \widehat{K^2_{RG}}) = \widehat{K^1_{NH}}$;
- $B^2_{NH}(\widehat{K^1_{RG}}, \widehat{K^2_{RG}}) = \widehat{K^2_{NH}}$.

Similarly as before, throughput of patients and utilisation is calculated for each cut-off quadruple $(K^1_{NH}, K^2_{NH}, K^1_{RG}, K^2_{RG})$. Optimal throughput, the Nash throughput and the PoA are also obtained.

The model is very computationally expensive; for the base scenario with $c_{NH} = 8, c_{RG} = 16$, each CCU has to consider 6885 strategies, since

$$\begin{aligned}
&\sum_{K^1_{NH}=0}^{c_{NH}} \sum_{K^2_{NH}=K^1_{NH}}^{c_{NH}} \sum_{K^1_{RG}=0}^{c_{RG}} \sum_{K^2_{RG}=K^1_{RG}}^{c_{RG}} (K^1_{NH}, K^2_{NH}, K^1_{RG}, K^2_{RG}) \\
&= \left(\sum_{K^1_{NH}=0}^{c_{NH}} (K^1_{NH}+1)\right) \times \left(\sum_{K^1_{RG}=0}^{c_{RG}} (K^1_{RG}+1)\right) \\
&= \frac{(c_{NH}+1)(c_{NH}+2)}{2} \times \frac{(c_{RG}+1)(c_{RG}+2)}{2}
\end{aligned}$$

For comparison, in Model 1 and Model 2, each CCU had to consider $(c_{NH}+1) \times (c_{RG}+1) = 153$ strategies. Due to the amount of strategies possible, the program is written in the programming language *Python*. The code is parallelised dividing the exploration of the strategy space into parallel jobs and is run on Merlin. Merlin is Cardiff University's High Performance Computer, some of its specifications include ([124]):

- consists of 2048 cores Intel Sandy Bridge processors (2.6GHz / 4GB per core / 8 cores per processor);
- an additional 864 cores Intel Westmere (2.8GHz / 3GB per core / 6 cores per processor) as a serial / high throughput subsystem;
- configured with 8+TB of total memory across the entire cluster, with a 50 TB global parallel file storage.

The next two sections will describe the modified Markov chains. In Section 8.3, Model 3 will be introduced, where total rejections are possible and in Section 8.4 Model 4 will be presented where total rejections are not allowed.

8.3 Model 3

This model assumes that if the bed occupancy level at both Units exceeds a predetermined transfer cut-off, admissions to CCU are cancelled and patients will be admitted to a Ward within the hospital (as in Model 1). The underlying Markov chain for a network comprising two CCUs is presented in Figure 8.1.

The overall arrival rates ($\widetilde{\lambda}_1$ at NH and $\widetilde{\lambda}_2$ at RG) at each CCU are as follows:

$$\widetilde{\lambda}_1 = \begin{cases} \lambda_{1a}, & \text{if } i < K^1_{NH} \text{ and } j < K^2_{RG} \\ \lambda_{1b}, & \text{if } K^1_{NH} \leq i < K^2_{NH} \text{ and } j < K^2_{RG} \\ \lambda_{1a} + \lambda_{2b}, & \text{if } i < K^1_{NH} \text{ and } j \geq K^2_{RG} \\ \lambda_{1b} + \lambda_{2b}, & \text{if } K^1_{NH} \leq i < K^2_{NH} \text{ and } j \geq K^2_{RG} \\ 0, & \text{if } i \geq K^2_{NH} \end{cases}$$

$$\widetilde{\lambda}_2 = \begin{cases} \lambda_{2a}, & \text{if } i < K^2_{NH} \text{ and } j < K^1_{RG} \\ \lambda_{2b}, & \text{if } i < K^2_{NH} \text{ and } K^1_{RG} \leq j < K^2_{RG} \\ \lambda_{2a} + \lambda_{1b}, & \text{if } i \geq K^2_{NH} \text{ and } j < K^1_{RG} \\ \lambda_{2b} + \lambda_{1b}, & \text{if } i \geq K^2_{NH} \text{ and } K^1_{RG} \leq j < K^2_{RG} \\ 0, & \text{if } j \geq K^2_{RG} \end{cases}$$

Figure 8.2 illustrates the overall arrival rates for each CCU, $(\widetilde{\lambda}_1, \widetilde{\lambda}_2)$ in each region separated by the cut-off points.

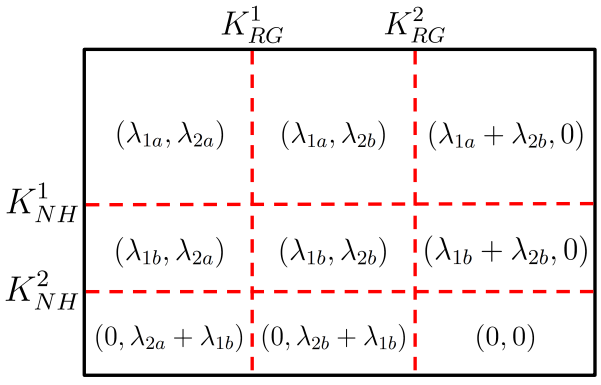


Figure 8.2: Arrival rate parameters for each CCU in each region

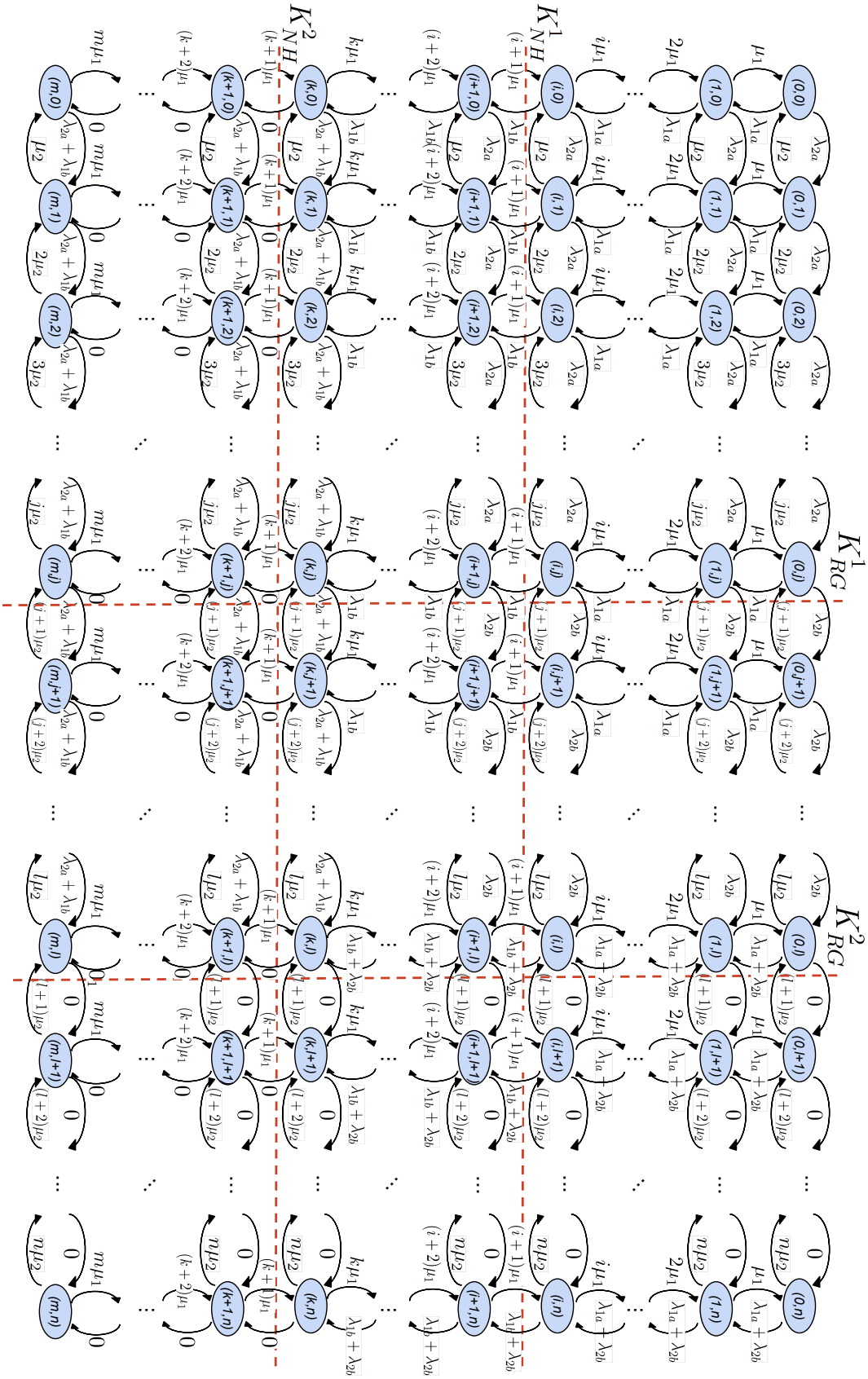


Figure 8.1: Markov chain for Model 3

Therefore, the transition matrix Q is obtained for the following transition rates $q_{i,j}$:

$$q_{i,j} = \begin{cases} u_i\mu_1 & \text{if } (u_i, v_i) - (u_j, v_j) = (1, 0), \\ v_i\mu_2 & \text{if } (u_i, v_i) - (u_j, v_j) = (0, 1), \\ \lambda_{1a} & \text{if } (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } u_i < K^1_{NH} \text{ and } v_i < K^2_{RG}, \\ \lambda_{1b} & \text{if } (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } K^1_{NH} \leq u_i < K^2_{NH} \text{ and } v_i < K^2_{RG}, \\ \lambda_{2a} & \text{if } (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } u_i < K^2_{NH} \text{ and } v_i < K^1_{RG}, \\ \lambda_{2b} & \text{if } (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } u_i < K^2_{NH} \text{ and } K^1_{RG} \leq v_i < K^2_{RG}, \\ \lambda_{1a} + \lambda_{2b} & \text{if } (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } u_i < K^1_{NH} \text{ and } v_i \geq K^2_{RG}, \\ \lambda_{1b} + \lambda_{2b} & \text{if } (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } K^1_{NH} \leq u_i < K^2_{NH} \text{ and } v_i \geq K^2_{RG}, \\ \lambda_{2a} + \lambda_{1b} & \text{if } (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } u_i \geq K^2_{NH} \text{ and } v_i < K^1_{RG}, \\ \lambda_{2b} + \lambda_{1b} & \text{if } (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } u_i \geq K^2_{NH} \text{ and } K^1_{RG} \leq v_i < K^2_{RG}, \\ 0 & \text{otherwise.} \end{cases}$$

The same set of parameters as in Model 1 and 2, obtained from the data, will be used:

Table 8.1: Parameter values used in the model

Parameter	**Parameter description**	**Parameter value**
m	the bed capacity at NH	8
n	the bed capacity at RG	16
λ_{1a}	the arrival rate at NH	1.498
λ_{2a}	the arrival rate at RG	2.245
μ_1	the service rate at NH	0.262
μ_2	the service rate at RG	0.198
target	bed utilisation target	80%

For example, for a given set of parameters, no increase in demand and a reduction rate after the first cut-off of 10%, so that $\lambda_{1b} = 0.9 \times \lambda_{1a}$ and $\lambda_{2b} = 0.9 \times \lambda_{2a}$, and target between 0.6 and 0.9, the PoA is investigated, and results are presented in Table 8.2.

For 80% target the Nash equilibrium and also the optimum throughput of 3.63 is at $(8, 8, 16, 16)$, hence the PoA=1. Having $K^1_{NH} = K^2_{NH} = c_{NH}$ and $K^1_{RG} = K^2_{RG} = c_{RG}$ means that for the current demand both Units can cope with the demand, so there is no need to decrease the arrival rate after the first cut-off at both CCUs, and both CCU will be running at a utilisation rate of less than 80%. If the target is lowered to 70%, Nash equilibrium is at $(8, 8, 12, 16)$, meaning that NH would not have to decrease or transfer patients; however, when the bed occupancy at RG is 12, arrival

Table 8.2: Model 3 results for reduction rate of 10% and no increase demand

Target	**Nash equilibrium**	**Nash overall throughput**	**Optimal throughput**	**PoA**
0.6	(6,6,11,11)	3.07	3.63	1.18
0.7	(8,8,12,16)	3.56	3.62	1.02
0.8	(8,8,16,16)	3.63	3.63	1
0.9	(8,8,16,16)	3.63	3.63	1

rate would be decreased by 10%, resulting in throughput at Nash of 3.56. The optimal throughput is 3.62, giving a PoA of 1.02, meaning that 2% of all patients were turned away in a reduction stage.

Furthermore, the PoA will be evaluated for cases where the altered variables will be: target, demand increase and reduction rate.

8.3.1 'What if': Target, Demand and Reduction Rate Change

This section will investigate the effect on the PoA of changing the utilisation target and certain parameters. The PoA will be calculated for:

- target $\in [0.1, 1]$ in 0.1 steps;
- demand increase, $d \in [0, 2]$ in 0.1 steps, hence $\widehat{\lambda_{1a}} = (1+d) \times \lambda_{1a}$ and $\widehat{\lambda_{2a}} = (1+d) \times \lambda_{2a}$;
- reduction rate, $r \in [0.1, 1]$ in 0.1 steps, hence $\widehat{\lambda_{1b}} = (1-r) \times \widehat{\lambda_{1a}}$ and $\widehat{\lambda_{2b}} = (1-r) \times \widehat{\lambda_{2a}}$.

As an illustration, a target of 80% is chosen to illustrate how the PoA is affected by different demand increase and reduction rates, which is presented in Figure 8.3.

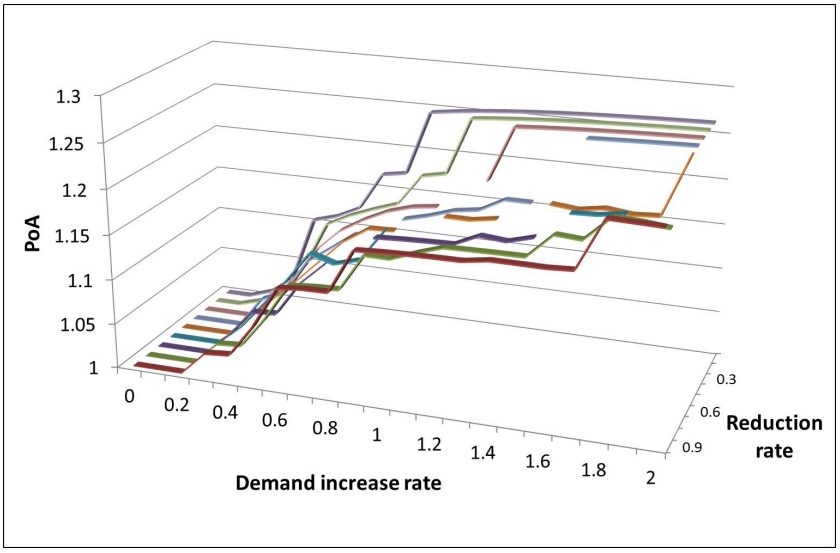


Figure 8.3: The PoA for target of 80% for Model 3

Visibly, there are some gaps for certain values; in those gaps the Nash equilibrium does not exist in pure strategies (as discussed in Chapter 7).

In general, the PoA increases with demand increase (as in Model 1). For example, for a reduction rate, $r = 0.1$ and demand rate, d between 0.1 and 0.6, the PoA is investigated, and results are presented in Table 8.3.

Table 8.3: Model 3 results for target of 80% and reduction rate of 0.1

Demand increase	**Nash equilibrium**	**Nash overall throughput**	**Optimal throughput**	**PoA**
0.1	(8,8,16,16)	3.89	3.89	1
0.2	(8,8,14,16)	4.08	4.11	1.01
0.3	(8,8,14,15)	4.17	4.29	1.03
0.4	(7,7,14,14)	4.02	4.43	1.10
0.5	(7,7,14,14)	4.09	4.55	1.11
0.6	(7,7,12,14)	4.12	4.64	1.13

For a 0.1 demand increase, CCUs do not compete, as both never reduce the number of patients coming in, or never transfer any patients; hence no inefficiency is observed and the PoA=1. As the demand increases to 0.2, a little inefficiency can be observed, RG reduces their arrival rate when there are 14 beds occupied. Inspection of RG utilisation rates for different RG cut-off points showed that if RG picked $K^1_{RG} = 16$ and $K^2_{RG} = 16$, the utilisation would be 80.71%, hence above the 80% target. If K^1_{RG} was reduced to 15, the utilisation would be 80.41%, which is still too high. For $K^1_{RG} = 14$, the utilisation rate is just below 80%, i.e. 79.93%. Therefore the Nash is at $(8, 8, 14, 16)$. For 0.3 demand increase, RG starts to transfer some patients over to NH. As demand increases further, NH after receiving patients from RG now also declares transfer status, and RG reduction status starts at an even lower bed occupancy.

Note that for 0.2 demand increase, RG preferred to reduce arrivals and hence loose a proportion of patients instead of reducing K^2_{RG} to 15. In that case, some RG patients would have been transferred to NH and no patients would be lost. For the $(8, 8, 15, 15)$ strategy profile, the utilisation rate at NH would be 76.85% and at RG 77.36%. Since RG is selfish and tries to get the utilisation rate as close to 80%, RG would move the strategy profile to $(8, 8, 14, 16)$, because the utilisation rate for RG for this strategy would be 79.93%, as before. In a centralised system, the strategy profile $(8, 8, 15, 15)$ would be preferred.

In general, the PoA decreases as the target value increases. As an illustration, for a reduction rate, r, of 0.3 and target values between 0.2 and 1, the PoA is illustrated in Figure 8.4. The PoA for a target value of 0.1 is very high (between 19.4484 for a 0 demand increase and 26.1796 for a increase demand of 2) and therefore is excluded from the graph.

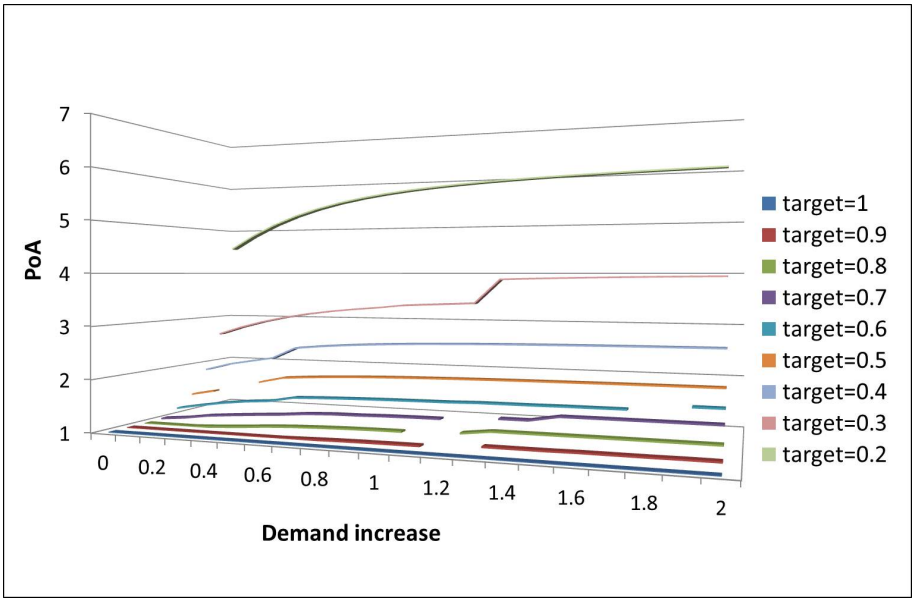


Figure 8.4: The PoA for different target values and reduction rate of 0.3 for Model 3

As before, for the cases, where the PoA$\rightarrow \infty$, gaps in the graph occur. Visibly, for each target value, the PoA increases with demand. For 100% target, the PoA is always 1, even for a very high demand. It was also observed in Model 1, and the reason is the same, setting a 100% target implies that both CCUs will try to maximise throughput.

The next model will inspect the effect of not allowing rejections in the transfer status on the PoA.

8.4 Model 4

This model assumes that if the bed occupancy levels at both CCUs exceed a pre-determined second cut-off level, then transfers are not allowed and each CCU has to accommodate their own patients. The Markov chain used to model this game is shown in Figure 8.5.

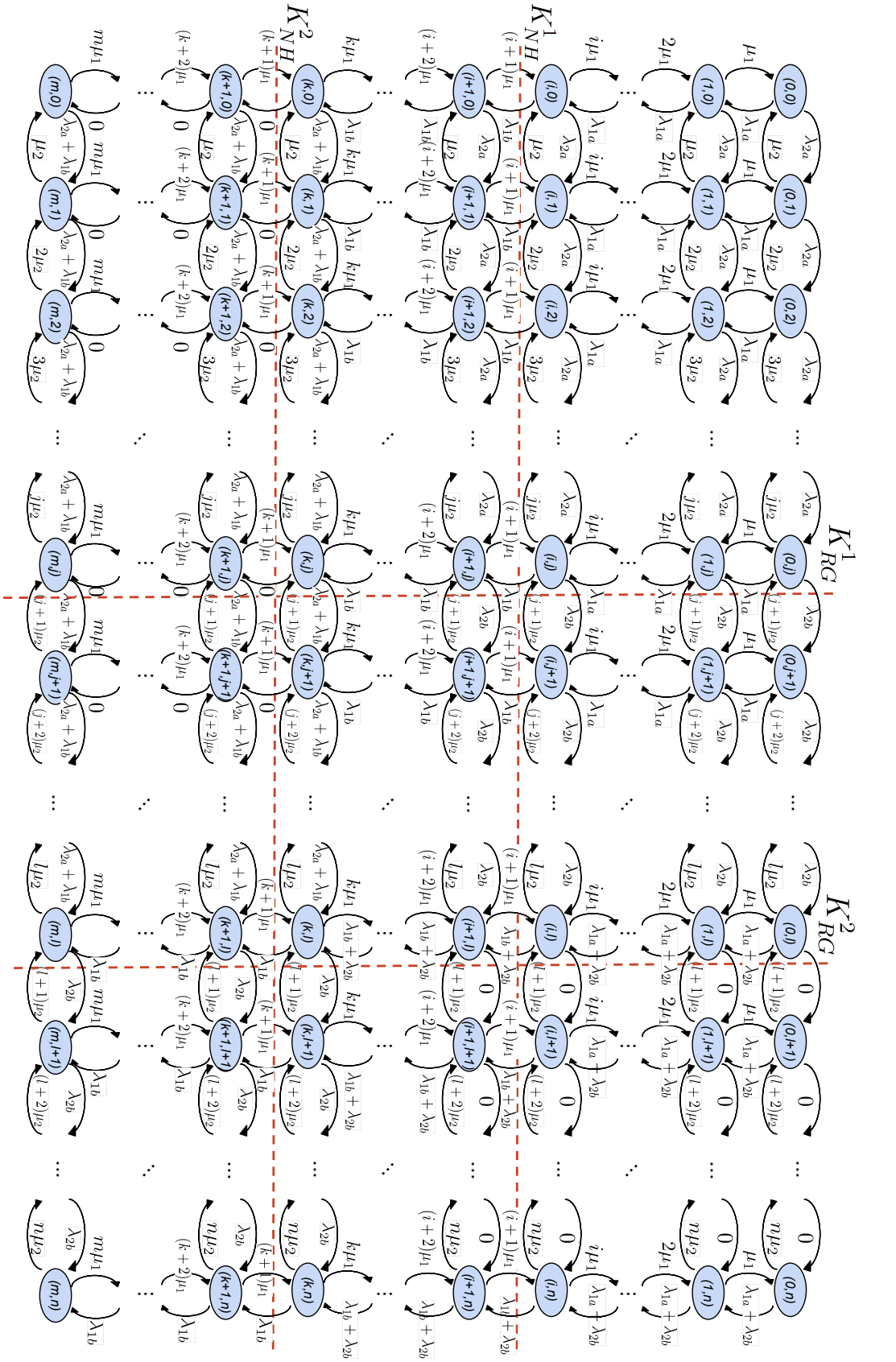


Figure 8.5: Markov chain for Model 4

The overall arrival rates ($\widetilde{\lambda}_1$ at NH and $\widetilde{\lambda}_2$ at RG) at each CCU are as follows:

$$\widetilde{\lambda}_1 = \begin{cases} \lambda_{1a}, & \text{if } u_i < K^1_{NH} \text{ and } v_i < K^2_{RG} \\ \lambda_{1b} & \text{if } \begin{cases} K^1_{NH} \le u_i < K^2_{NH} \text{ and } v_i < K^2_{RG} \text{ or} \\ u_i \ge K^2_{NH} \text{ and } v_i \ge K^2_{RG}, \end{cases} \\ \lambda_{1a} + \lambda_{2b}, & \text{if } u_i < K^1_{NH} \text{ and } v_i \ge K^2_{RG} \\ \lambda_{1b} + \lambda_{2b}, & \text{if } K^1_{NH} \le u_i < K^2_{NH} \text{ and } v_i \ge K^2_{RG} \\ 0, & \text{if } i \ge K^2_{NH} \text{ and } v_i < K^2_{RG} \end{cases}$$

$$\widetilde{\lambda}_2 = \begin{cases} \lambda_{2a}, & \text{if } u_i < K^2_{NH} \text{ and } v_i < K^1_{RG} \\ \lambda_{2b}, & \text{if } \begin{cases} u_i < K^2_{NH} \text{ and } K^1_{RG} \le v_i < K^2_{RG} \text{ or} \\ u_i \ge K^2_{NH} \text{ and } v_i \ge K^2_{RG} \end{cases} \\ \lambda_{2a} + \lambda_{1b}, & \text{if } u_i \ge K^2_{NH} \text{ and } v_i < K^1_{RG}, \\ \lambda_{2b} + \lambda_{1b}, & \text{if } u_i \ge K^2_{NH} \text{ and } K^1_{RG} \le v_i < K^2_{RG} \\ 0, & \text{if } i \le K^2_{NH} \text{ and } j \ge K^2_{RG} \end{cases}$$

Figure 8.6 illustrates the overall arrival rates for each CCU, $(\widetilde{\lambda}_1, \widetilde{\lambda}_2)$ in each region separated by the cut-off points.

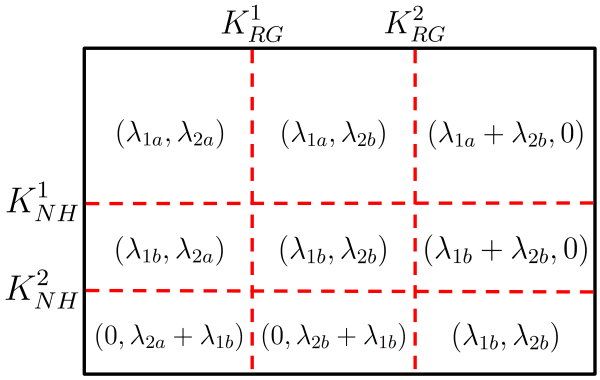


Figure 8.6: Arrival rate parameters for each CCU in each region

Therefore, the transition matrix Q is obtained for the following transition rates $q_{i,j}$:

$$q_{i,j} = \begin{cases} u_i\mu_1 & \text{if } (u_i, v_i) - (u_j, v_j) = (1, 0), \\ v_i\mu_2 & \text{if } (u_i, v_i) - (u_j, v_j) = (0, 1), \\ \lambda_{1a} & \text{if } (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } u_i < K^1_{NH} \text{ and } v_i < K^2_{RG}, \\ \lambda_{1b} & \text{if } (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } \begin{cases} K^1_{NH} \leq u_i < K^2_{NH} \text{ and } v_i < K^2_{RG} \text{ or} \\ u_i \geq K^2_{NH} \text{ and } v_i \geq K^2_{RG}, \end{cases} \\ \lambda_{2a} & \text{if } (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } u_i < K^2_{NH} \text{ and } v_i < K^1_{RG}, \\ \lambda_{2b} & \text{if } (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } \begin{cases} u_i < K^2_{NH} \text{ and } K^1_{RG} \leq v_i < K^2_{RG} \text{ or} \\ u_i \geq K^2_{NH} \text{ and } v_i \geq K^2_{RG}, \end{cases} \\ \lambda_{1a} + \lambda_{2b} & \text{if } (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } u_i < K^1_{NH} \text{ and } v_i \geq K^2_{RG}, \\ \lambda_{1b} + \lambda_{2b} & \text{if } (u_i, v_i) - (u_j, v_j) = (-1, 0) \text{ and } K^1_{NH} \leq u_i < K^2_{NH} \text{ and } v_i \geq K^2_{RG}, \\ \lambda_{2a} + \lambda_{1b} & \text{if } (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } u_i \geq K^2_{NH} \text{ and } v_i < K^1_{RG}, \\ \lambda_{2b} + \lambda_{1b} & \text{if } (u_i, v_i) - (u_j, v_j) = (0, -1) \text{ and } u_i \geq K^2_{NH} \text{ and } K^1_{RG} \leq v_i < K^2_{RG}, \\ 0 & \text{otherwise.} \end{cases}$$

The same set of parameters as in Model 3 are used. Assume a reduction rate of 10% is applied, so that $\lambda_{1b} = 0.9 \times \lambda_{1a}$ and $\lambda_{2b} = 0.9 \times \lambda_{2a}$. The best responses are obtained; the Nash throughput and also the optimum throughput is 3.6255 and is at $(8, 8, 16, 16)$, hence the PoA=1. For these parameters the results are the same as in Model 3, since the CCUs are not transferring patients due to the choice of the second cut-off point. However, if the target is lowered to 70%, the Nash equilibrium is at $(8, 8, 15, 15)$, meaning that a proportion of patients will be transferred to NH. The optimal throughput is 3.6255 and the Nash throughput is 3.6075, resulting in a PoA of 1.005, which means that 0.5% of patients are rejected.

Furthermore, similarly to Model 3, the PoA will be evaluated for cases where the altered variables will be as before: target, demand increase and reduction rate.

8.4.1 'What if': Target, Demand and Decrease Rate Change

This section will investigate the effect on the PoA of changing the utilisation target and certain parameters. The PoA will be calculated for:

- target $\in [0.1, 1]$ in 0.1 steps;
- demand increase, $d \in [0, 2]$ in 0.1 steps, hence $\widehat{\lambda_{1a}} = (1 + d) \times \lambda_{1a}$ and $\widehat{\lambda_{2a}} = (1 + d) \times \lambda_{2a}$;
- reduction rate, $r \in [0.1, 1]$ in 0.1 steps, hence $\widehat{\lambda_{1b}} = (1 - r) \times \widehat{\lambda_{1a}}$ and $\widehat{\lambda_{2b}} = (1 - r) \times \widehat{\lambda_{2a}}$.

As an illustration, a target of 80% is chosen to illustrate how the PoA is affected by different demand increase and reduction rates, which is presented in Figure 8.7.

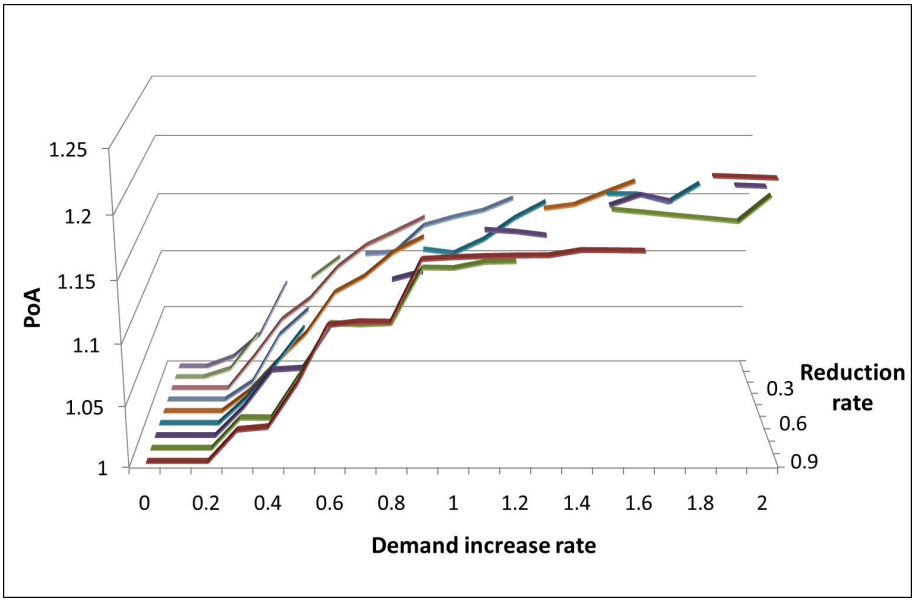


Figure 8.7: The PoA for a target of 80% for Model 4

Visibly, there are even more gaps than in Model 3; and in fact as the target decreases, the number of cases for which pure Nash equilibria do not exist increases.

In general, the PoA increases with demand increase. For example, for a reduction rate, $r = 0.3$ and demand rate, d, between 0.2 and 0.7, the PoA is investigated, and results are presented in Table 8.4.

Table 8.4: Model 4 results for target of 80% and reduction rate of 0.3

Demand increase	**Nash equilibrium**	**Nash overall throughput**	**Optimal throughput**	**PoA**
0.2	(8,8,16,16)	4.07	4.07	1
0.3	(8,8,14,15)	4.12	4.25	1.03
0.4	(7,8,13,15)	4.13	4.39	1.06
0.5	(4,8,14,14)	4.17	4.50	1.08
0.6	(6,8,9,16)	4.15	4.59	1.11
0.7	(5,6,11,12)	4.15	4.67	1.13

For a demand increase of 0.2, no inefficiency is observed; both CCUs never reduce or transfer patients. As the demand is increased to 0.3, RG starts reducing and transferring patients, resulting in a PoA of 1.03; hence 3.00% of patients are lost. As the demand increases even further, both cut-off points for both CCUs decrease, reflecting in an increase of the PoA.

In general, the PoA decreases with target value. As an illustration, for a reduction rate, r, of 0.9 and target values between 0.2 and 1 the PoA is illustrated in Figure 8.8. For a target value of 0.1 the PoA$\rightarrow \infty$ mainly, and therefore is excluded from the graph.

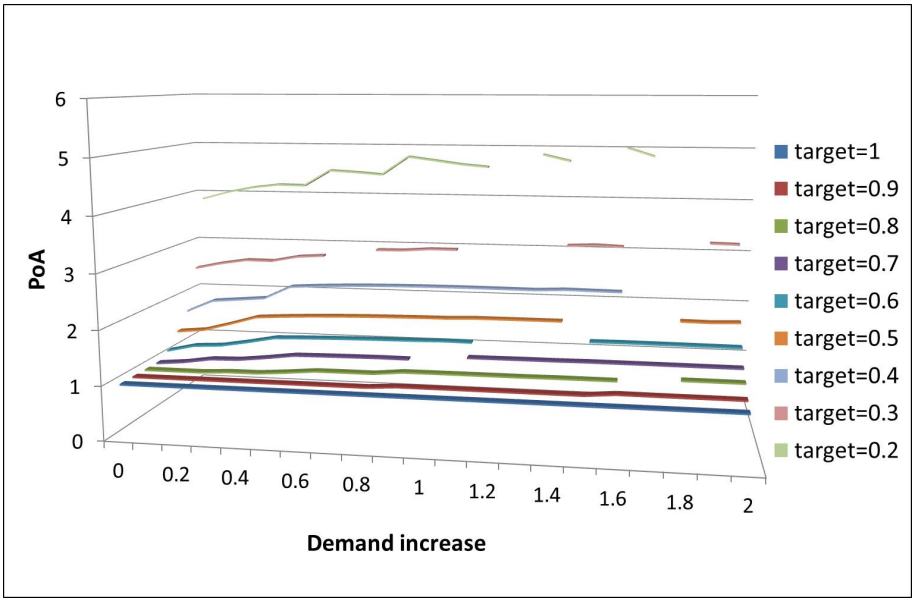


Figure 8.8: The PoA for different target values and reduction rate of 0.9 for Model 4

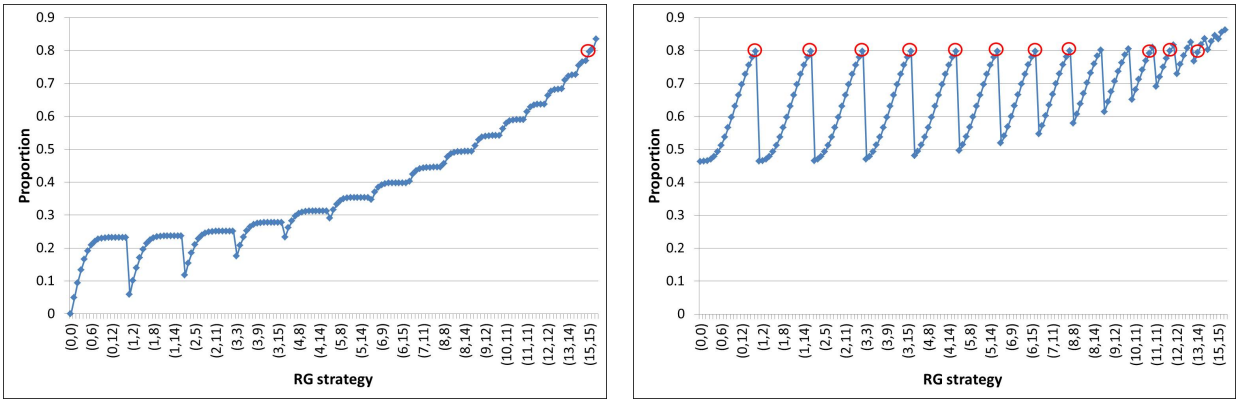
In general, the PoA increases with demand; however, in some cases, for example for a target of 0.2, the PoA slightly decreases as demand increases. The reason is the same as described in Section 7.8.2. The Nash equilibria stay the same and since increase of demand, optimal throughput increases; therefore the PoA decreases.

8.5 Conclusions

To formally investigate the impact of decentralised decision making, where each player has a choice of two strategy points to pick, queueing network models have been developed. Two models have been considered; one where total rejections after the transfer cut-off were allowed and the second where there were not allowed. The inefficiency which occurred has been measured and presented in a graphical form.

In general, in Model 4, for a lot of cases, especially for low target values, the Nash equilibrium does not exist. It is due to the amount of feasible strategies; a very similar utilisation rate could be obtained for a lot of different cut-off pairs, which results in non stationary behaviour from the players (analogous to the matching pennies game shown in Figure 7.1). As an illustration assume that NH picked $(8, 8)$ as their strategy for an 80% target, a 0.4 demand increase and a reduction rate of 0.2, the utilisation rate at RG for various strategies are shown in Figures 8.9a and 8.9b.

For Model 4 it can be seen that for 11 different strategies, the utilisation rate was between 79% and 80%, while for Model 3 this was true for only 1 strategy, due to the fact that no rejections are allowed (ensuring high utilisation rates). To further study this model the restrictions of pure strategies could be relaxed.



(a) Model 3

(b) Model 4

Figure 8.9: Utilisation rates for NH strategy (8,8), an 80% target, a 0.4 demand increase and a reduction rate of 0.2

The work presented in the last two chapters has attempted to quantify the effect of selfish behaviour of CCUs. The main message to the hospital managers would be that their selfish behaviour can potentially decrease effectiveness of the whole system, resulting in a lower overall performance, and that they should always consider the effect of their decisions on the whole system. With ever increasing demand, the inefficiency of the system due to selfish behaviour is expected to increase, but only up to the point, where the system is simply unable to cope and selfish behaviour does not have a negative effect. It has been shown that targets can be used by policy makers to ensure low levels of inefficiency; however, these must be chosen carefully, as they can also have a negative effect.

Chapter 9

Final Conclusions and Further Work

This chapter summarises some of the interesting and important findings of this thesis. The research included in this thesis can be divided into three main parts as shown in Figure 1.3. The first part, which includes the work presented in Chapters 2, 3 and 4, carries out an analysis of data provided by the University Hospital of Wales in Cardiff (UHW). Both computer simulation and analytical queueing models were developed and various 'what if' scenarios considered. The second part, which includes the work presented in Chapters 5 and 6, described the work undertaken with managers from the Royal Gwent and the Nevill Hall Hospitals. Data from both CCUs were analysed and a theoretical queueing model was considered. In the final part of the thesis, which includes the work presented in Chapters 7 and 8, game theoretical models were presented. Each of the parts will be now reviewed, the models and results will be briefly discussed, and finally an indication of future research where appropriate will be given.

9.1 Part I

9.1.1 Summary of Chapters 2, 3 and 4

In that this part appropriate statistical distributions of inter-arrival times and lengths of stay were determined. Very importantly, it was shown that any analysis should account for different patient types, namely elective and emergency, since arrival and service patterns are different for both category patients.

Two main objectives of this part of the thesis were to suggest measures which could be implemented to increase the throughput and to determine ways in which the degree of variation in bed occupancy could be reduced. Both simulation modelling and analytical queueing techniques were utilised to address both objectives.

The simulation model which was built in Visual Basic replicated the bed occupancy levels accu-

rately. The effect of implementing some new policies regarding cancellations and/or increasing elective surgeries was examined. It was shown that by allowing extra elective admissions (up to 4 per day) when the Unit is running at a relatively low capacity, and not allowing any elective admissions when the bed occupancy exceeded a predetermined cut-off level of 24, the standard deviation of bed occupancy was reduced by 17.8% and the throughput of patients was increased by 4.8%.

A $M_2/M_2/c/c/FIFO$ queueing model with random arrivals from two different streams and corresponding negative exponential service times depending on patients' type was developed. The differential-difference equations to describe the system were set up and solved. The formula for the probability of having i emergency and j elective patients present in the system at steady state was obtained.

Further to this, a $M_2/M/c/c+m/FIFO$ has been considered with two separate arrival streams of patients and a combined service rate, where a queue is allowed to form. The differential-difference equations to describe the system were set up and solved. The formula for the probability that there are i emergency and j elective patients present with different arrival rates and combined service rate μ in the system where a queue of size m is allowed was obtained.

A connection between $M_2/M_2/c/c/FIFO$ and $M/M/c/c/FIFO$ was established. It was shown that the probability of having n patients in the system, P_n, can be obtained from the probability of having i emergency and j elective patients present in the system, $P_{i,j}$. Using a combinatorial argument it was shown that $P_{i,j}$ can be obtained from P_n. More importantly, the probability of having $j_i \in \mathbb{Z}$ customers of type i in an $M_k/M_k/c/c/FIFO$ queue having an arrival rate λ_i and a service rate μ_i for $i \in [k]$ was obtained using similar argument as for the two customer types.

Finally, two expressions were developed, one for the bed occupancy probabilities, with a restriction on the number of customers allowed in the system, with a cut-off point k, and the second for the bed occupancy probabilities with a cut-off point and an increase in elective admissions for the bed occupancy between predetermined levels a and b.

The final chapter of this part of the thesis investigated further applications of mathematical modelling. Since elective demand is very much dependent on the day of the week, time-dependent aspects of bed probabilities were considered using the developed mathematical models and Euler's numerical method for solving differential equations. Information obtained from the mathematical model of expected levels of bed occupancies might be useful to the Director of the CCU to advise decision making regarding admission of extra elective patients in the near future. The final section in this part of the thesis provided information regarding the number of nurses required on each day of the week using two different approaches, one that optimises an actual expected cost and the

second that optimises the expected wastage cost, using a classic model from inventory theory: the Newsboy model. Both approaches, for two different cases where the nurse to patient ratio is 1:1 and where the nurse to patient ratio is variable, recommended the same number of nurses.

The queueing models described have not been developed in the literature and thus are considered to be original research contributions offered by this thesis. The work from this part of the thesis has been published in Griffiths *et al.*, 2013 [74].

9.1.2 Limitations and Further Work

The reason for imperfect bed occupancy model could have been caused by long length of stay tails. Effectively, there exist a third group of arrivals (patients in vegetative state). Arrival rate is small and difficult to estimate, length of stay in extreme and their length of stay virtually impossible to estimate. Also, in Section 4.4 assumption was made regarding nurse to patients ratios. In real life they depend on the number of Level 3, Level 2 and Level 1 patients; however, the date was not available with this information therefore as a proxy it is assumed that 1 emergency patient require 1 nurse, and 2 elective patients require 1 nurse.

A further possibility would be to investigate the effect of the CCU being able to "ring fence" some of their beds for elective admissions. Currently, elective patients who require a stay in the CCU after their surgery, must first be given the green light by the CCU before their surgery can take place. This ensures that there is adequate provision for them. The possibility of ring fencing beds could reduce the cancellations of elective patients; however this might increase transfers of emergency patients to other CCUs.

An extension to the work presented in Section 3.3.4 was considered for the queue $M_2/M/c/c+m/FIFO$. Recall, a queueing model is considered where a queue of size m is allowed. Let $P_{i,j,\tilde{i},\tilde{j}}$ be the probability of having:

- i emergency patients in the service
- j elective patients in the service
- $\tilde{i}$ emergency patients in the queue
- $\tilde{j}$ elective patients in the queue

Therefore $i + j = c$, $\tilde{i} + \tilde{j} = m$, $I = i + \tilde{i}$ and $J = j + \tilde{j}$, where I is the number of emergency patients in the system, and J is the number of elective patients in the system.

Following the proof in Section 3.3.4,

$$P_{i,j,\tilde{i},\tilde{j}|i+j+\tilde{i}+\tilde{j}} = \frac{P_{(i,j,\tilde{i},\tilde{j})\cap(i+j+\tilde{i}+\tilde{j})}}{P_{i+j+\tilde{i}+\tilde{j}}} = \frac{P_{i,j,\tilde{i},\tilde{j}}}{P_{i+j+\tilde{i}+\tilde{j}}}$$

Therefore:

$$P_{i,j,\tilde{i},\tilde{j}} = P_{i,j,\tilde{i},\tilde{j}|i+j+\tilde{i}+\tilde{j}} \times P_{i+j+\tilde{i}+\tilde{j}}$$

where $P_{i+j+\tilde{i}+\tilde{j}} = P_{c+m}$ and the results are known and given in Stewart, 2009 [152]. $P_{i,j,\tilde{i},\tilde{j}|i+j+\tilde{i}+\tilde{j}}$ is the probability of having i emergency and j elective patients in the service, $\tilde{i}$ emergency and $\tilde{j}$ elective patients in the queue given $i+j+\tilde{i}+\tilde{j}$ patients in the system. Thus:

$$P_{i,j,\tilde{i},\tilde{j}|i+j+\tilde{i}+\tilde{j}} = \binom{i+j}{i} p^i q^j \times \binom{\tilde{i}+\tilde{j}}{\tilde{i}} \tilde{p}^{\tilde{i}} \tilde{q}^{\tilde{j}}$$

where:

$p = \dfrac{\frac{\lambda_1}{\mu_1}}{\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}}$ probability that a server is used by an emergency patient

$q = \dfrac{\frac{\lambda_2}{\mu_2}}{\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}}$ probability that a server is used by an elective patient

$\tilde{p} = \dfrac{\lambda_1}{\lambda_1 + \lambda_2}$ probability that there is an emergency patient in the queue

$\tilde{q} = \dfrac{\lambda_2}{\lambda_1 + \lambda_2}$ probability that there is an elective patient in the queue

The case $i+j \leq c$ has been already considered, the only interest is now in the case when $I + J > c$. Since $max(0, c-J) \leq i \leq min(c, I)$,

$$P_{I,J|I+J} = \sum_{i=max(0,c-J)}^{min(c,I)} \binom{i+j}{i} p^i q^j \times \binom{I+J-c}{I-i} \tilde{p}^{I-i} \tilde{q}^{J-c+i}$$

The summation proved to be quite complicated, but obtaining a concise formula for this summation could lead to an elegant proof of the formula for $P_{i,j,\tilde{i},\tilde{j}}$.

9.2 Part II

9.2.1 Summary of Chapters 5 and 6

This part of the thesis described the project undertaken with CCU managers from the two hospitals from the Aneurin Bevan Local Health Board. Initially, data from both CCUs was analysed to determine arrival and service patterns. Recall that the bed capacity in the Royal Gwent was changed

twice, resulting in a novel consideration of each period separately and then combining them using the given weights, dependent on the duration of each period. Interesting findings were made regarding the bed capacity change; it did not influence a change in arrival rate, but increased the average duration of stay.

Moreover, a state-dependent queueing model was developed, where the admission rates are dependent of the current bed occupancy level. A system with c service channels and cut-off points k_1 and k_2 was described by a set of differential-difference equations, which then were solved to get the bed occupancy probabilities. Since the cut-off points and corresponding arrival rates were impossible to obtain from the data sets, an optimisation problem was defined to obtain these parameters. The model provided an accurate representation of the data for both CCUs, and a few 'what if' scenarios were considered. The first considered the transfer of patients between the CCUs, and the second the transfer of patients along with proportion of beds from one CCU to the other. It has been shown that sharing resources between both CCUs is beneficial not only for CCUs but also for patients.

The main 'what if' scenario was developed as a result of an ongoing plan for the new Specialist Critical Care Centre (SCCC) to be built. The two CCUs were treated as one with a shared bed capacity. The model which was fitted to the combined Unit compares favourably with the observed values, and was used to test changes in Unit size. The main finding was that by increasing the bed capacity by two, to 25 beds, the combined Unit would run on a relatively high bed utilisation less than 30% of the time, with a low probability of rejection and a utilisation rate of 72.57%. Furthermore, the model was used to explore a scenario, which investigated the effect of increasing the demand or decreasing the average delay to discharge time.

9.2.2 Limitations and Further Work

The data set provided did not contain details about patients referred for critical care who were not admitted due to all beds being occupied. An unfortunate consequence is that using such arrival data in a queueing model of CCU does not correspond to the true referral rate.

In Chapter 5 bed blocking was considered: patients are considered to block a bed if they are well enough to leave the Unit, but for some reason remain, using up a valuable, expensive and limited resource. The model was then tested to consider changes in the average blocking time. Phase-type distributions could be examined, as they prove to be very useful when considering blocking, especially the two-phase Coxian distribution. The first stage is the 'actual' length of stay, the second is the blocking stage. Patients in the first stage are served with rate μ_1 and with rate μ_2 at the second stage. Thus, after receiving service at the first phase, the process, with probability α_1, continues on to the second phase, or with probability $1 - \alpha_1$, the service process is terminated and patients are

discharged without experiencing any delay. This is shown in Figure 9.1.

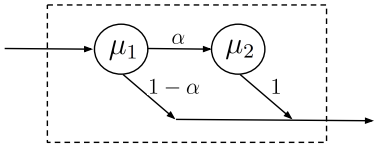


Figure 9.1: Two-phase Coxian distribution

9.3 Part III

9.3.1 Summary of Chapters 7 and 8

This part of the thesis was motivated by the fact that there were notable behavioural aspects, which were shown in Part II, in the running of the CCUs. Part III considered non-cooperative game theoretical models of the two CCUs, where both CCUs acted self-interestedly. The impact of the lack of cooperation was studied using the price of anarchy (PoA).

A review of the literature did not reveal any game theoretical work applied to the CCU environment; hence this work is considered to be an original research contribution.

Four models were considered. In the first two it was assumed that, if bed occupancy levels exceeded a pre-determined cut-off, the CCU managers requested transfer to the other CCU, provided there were available beds and the other CCU was not in 'transfer' status. Depending on the model, either patients were refused admission to a CCU (Model 1), or each CCU was forced to accept its own patients (Model 2).

Model 3 and 4 investigated a wider strategy space. It was assumed that if CCUs are running on relatively high bed occupancy levels they declared being in 'reduction' status first, where only a proportion of patients are allowed to be admitted, to later declare being in 'transfer' status as the bed occupancy increases.

The interaction was modelled through a two dimensional Markov chain, which was accordingly adapted to each model. The queueing network was then inserted within a static non-cooperative game, where each CCU manager was faced with an optimisation problem. A set of best responses for each CCU were obtained, to obtain the Nash equilibria in pure strategies. To measure the inefficiency created by the competitive interaction between CCUs, the price of anarchy (PoA) was used.

In Model 1 and 2, the PoA was calculated for a few scenarios where the target, demand and bed

capacity were altered. It was shown that in a CCU environment with inadequate bed capacity to provide the service, a high PoA is to be expected. In Model 3 and 4, the PoA was calculated for a scenario where the target, demand and reduction rate were altered.

An important message for hospital Directors is that in a system with the possibility of interaction, there is potential for inefficiency to occur due to selfish behaviour. They should always consider the impact on the whole system before any major decision is made, because as it was shown their selfish behaviour can indeed decrease effectiveness of the whole system.

9.3.2 Limitations and Further Work

For every model it was assumed that transferred patient obtain length of stay characteristic of the CCU that the patient is transferred to. A further model could be explored where patients carry length of stay characteristic from the CCU they were initially admitted to.

In Model 3 and 4, after the first cut-off point, it was assumed that only a proportion of patients was admitted, with the remaining percentage of patients being refused admission to a CCU. As an extension, a further two models could be considered, where the proportion of patients turned away were admitted to the other CCU, which could potentially decrease the PoA.

Another immediate extension to the work provided in Chapter 7 and 8 is to consider mixed Nash equilibria. It was stated previously that mixed equilibria are inadequate in the modelling of CCUs. However it would be very interesting to see how the PoA is affected from a theoretical point of view.

As a further extension, similar to the work in Roughgarden, 2005 [144], an analytical upper bound on the price of anarchy based on system parameters could be obtained.

9.4 Final Remarks

This thesis has considered many aspects of theoretical and practical applications of mathematical modelling in the CCU setting. The analysis has shown that by developing and applying queueing theory, improvements can be made in the management and efficiency of the CCUs. Although this thesis has used information and data from particular CCUs, the modelling described herein is of a generic nature, and hence could be applied with little amendment in other settings.

List of Figures

List of Tables

Appendix A

Proof of Theorem 3.3.2

Due to the form of $P_{i,j}$ various cases are considered. The required algebraic manipulation for each of the equations in 3.6 is given below:

(1) For $i = j = 0$:

$$\begin{aligned} P_{0,0} =& \frac{\mu P_{1,0} + \mu P_{0,1}}{\lambda_1 + \lambda_2} \\ =& \frac{\mu\theta_1 P_0 + \mu\theta_2 P_0}{\lambda_1 + \lambda_2} \\ =& \frac{\mu\frac{\lambda_1}{\mu} + \mu\frac{\lambda_2}{\mu}}{\lambda_1 + \lambda_2} P_0 \\ =& \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2} P_0 \\ =& P_0 \end{aligned}$$ □

(2) For $1 \leq i \leq (c-1)$ and $j = 0$:

$$\begin{aligned} P_{i,0} =& \frac{\lambda_1 P_{i-1,0} + (i+1)\mu P_{i+1,0} + (i+1)\mu P_{i,1}}{\lambda_1 + \lambda_2 + i\mu} \\ =& \frac{\lambda_1 \frac{1}{(i-1)!}{\theta_1}^{i-1} P_0 + (i+1)\mu\frac{1}{(i+1)!}{\theta_1}^{i+1} P_0 + (i+1)\mu\frac{1}{(i+1)!}{\theta_1}^{i}\theta_2 P_0}{\lambda_1 + \lambda_2 + i\mu} \\ =& \frac{\frac{1}{i!}{\theta_1}^{i}(\frac{i\lambda_1}{\theta_1} + \mu\theta_1 + \mu\theta_2) P_0}{\lambda_1 + \lambda_2 + i\mu} \\ =& \frac{\frac{1}{i!}{\theta_1}^{i}(i\mu + \lambda_1 + \lambda_2) P_0}{\lambda_1 + \lambda_2 + i\mu} \\ =& \frac{1}{i!}{\theta_1}^{i} P_0 \end{aligned}$$ □

(3) For $i = 0$ and $1 \leq j \leq (c-1)$:

$$\begin{aligned}
P_{0,j} =& \frac{\lambda_2 P_{0,j-1} + (j+1)\mu P_{0,j+1} + (j+1)\mu P_{1,j}}{\lambda_1 + \lambda_2 + j\mu} \\
=& \frac{\lambda_2 \frac{1}{(j-1)!}{\theta_2}^{j-1} P_0 + (j+1)\mu \frac{1}{(j+1)!}{\theta_2}^{j+1} P_0 + (j+1)\mu \frac{1}{(1+j)!}\theta_1 {\theta_2}^j P_0}{\lambda_1 + \lambda_2 + j\mu} \\
=& \frac{\frac{1}{j!}{\theta_2}^j (\frac{j\lambda_2}{\theta_2} + \mu\theta_2 + \mu\theta_1) P_0}{\lambda_1 + \lambda_2 + j\mu} \\
=& \frac{\frac{1}{j!}{\theta_2}^j (j\mu + \lambda_2 + \lambda_1) P_0}{\lambda_1 + \lambda_2 + j\mu} \\
=& \frac{1}{j!}{\theta_2}^j P_0
\end{aligned}$$

□

(4) For $i + j \leq (c-1)$ and $i, j \neq 0$:

$$\begin{aligned}
P_{i,j} =& \frac{\lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + (i+j+1)\mu P_{i+1,j} + (i+j+1)\mu P_{i,j+1}}{\lambda_1 + \lambda_2 + 2(i+j)\mu} \\
=& \frac{\frac{\lambda_1}{(i+j-1)!}{\theta_1}^{i-1}{\theta_2}^j + \frac{\lambda_2}{(i+j-1)!}{\theta_1}^i{\theta_2}^{j-1} + \frac{(i+j+1)\mu}{(i+j+1)!}{\theta_1}^{i+1}{\theta_2}^j + \frac{(i+j+1)\mu}{(i+j+1)!}{\theta_1}^i{\theta_2}^{j+1}}{\lambda_1 + \lambda_2 + 2(i+j)\mu} P_0 \\
=& \frac{\frac{1}{(i+j)!}{\theta_1}^i{\theta_2}^j \left(\frac{(i+j)\lambda_1}{\theta_1} + \frac{(i+j)\lambda_2}{\theta_2} + \mu\theta_1 + \mu\theta_2 \right)}{\lambda_1 + \lambda_2 + 2(i+j)\mu} P_0 \\
=& \frac{\frac{1}{(i+j)!}{\theta_1}^i{\theta_2}^j ((i+j)\mu + (i+j)\mu + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + 2(i+j)\mu} P_0 \\
=& \frac{\frac{1}{(i+j)!}{\theta_1}^i{\theta_2}^j (2(i+j)\mu + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + 2(i+j)\mu} P_0 \\
=& \frac{1}{(i+j)!}{\theta_1}^i{\theta_2}^j P_0
\end{aligned}$$

□

(5) For $i = 0$ and $j = c$:

$$\begin{aligned}
P_{0,c} =& \frac{\lambda_2 P_{0,c-1} + c\mu P_{1,c} + c\mu P_{0,c+1}}{\lambda_1 + \lambda_2 + c\mu} \\
=& \frac{\frac{\lambda_2}{(c-1)!}{\theta_2}^{c-1} + \frac{c\mu}{cc!}\theta_1{\theta_2}^c + \frac{c\mu}{cc!}{\theta_2}^{c+1}}{\lambda_1 + \lambda_2 + c\mu} P_0 \\
=& \frac{\frac{1}{c!}{\theta_2}^c \left(\frac{c\lambda_2}{\theta_2} + \mu\theta_1 + \mu\theta_2 \right)}{\lambda_1 + \lambda_2 + c\mu} P_0 \\
=& \frac{\frac{1}{c!}{\theta_2}^c (c\mu + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + c\mu} P_0 \\
=& \frac{1}{c!}{\theta_2}^c P_0
\end{aligned}$$

□

(6) For $i = c$ and $j = 0$:

$$\begin{aligned}
P_{c,0} =& \frac{\lambda_1 P_{c-1,0} + c\mu P_{c+1,0} + c\mu P_{c,1}}{\lambda_1 + \lambda_2 + c\mu} \\
=& \frac{\frac{\lambda_1}{(c-1)!}{\theta_1}^{c-1} + \frac{c\mu}{cc!}{\theta_1}^{c+1} + \frac{c\mu}{cc!}{\theta_1}^{c}\theta_2}{\lambda_1 + \lambda_2 + c\mu} P_0 \\
=& \frac{\frac{1}{c!}{\theta_1}^{c}\left(\frac{c\lambda_1}{\theta_1} + \mu\theta_1 + \mu\theta_2\right)}{\lambda_1 + \lambda_2 + c\mu} P_0 \\
=& \frac{\frac{1}{c!}{\theta_1}^{c}(c\mu + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + c\mu} P_0 \\
=& \frac{1}{c!}{\theta_1}^{c} P_0
\end{aligned}$$ □

(7) For $i + j = c$ and $i, j \neq 0$:

$$\begin{aligned}
P_{i,j} =& \frac{\lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + c\mu P_{i+1,j} + c\mu P_{i,j+1}}{\lambda_1 + \lambda_2 + 2c\mu} \\
=& \frac{\frac{\lambda_1}{(i+j-1)!}{\theta_1}^{i-1}{\theta_2}^{j} + \frac{\lambda_2}{(i+j-1)!}{\theta_1}^{i}{\theta_2}^{j-1} + \frac{c\mu}{cc!}{\theta_1}^{i+1}{\theta_2}^{j} + \frac{c\mu}{cc!}{\theta_1}^{i}{\theta_2}^{j+1}}{\lambda_1 + \lambda_2 + 2c\mu} P_0 \\
=& \frac{\frac{1}{(i+j)!}{\theta_1}^{i}{\theta_2}^{j}\left(\frac{(i+j)\lambda_1}{\theta_1} + \frac{(i+j)\lambda_2}{\theta_2} + \mu\theta_1 + \mu\theta_2\right)}{\lambda_1 + \lambda_2 + 2c\mu} P_0 \\
=& \frac{\frac{1}{c!}{\theta_1}^{i}{\theta_2}^{j}(c\mu + c\mu + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + 2c\mu} P_0 \\
=& \frac{\frac{1}{c!}{\theta_1}^{i}{\theta_2}^{j}(2c\mu + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + 2c\mu} P_0 \\
=& \frac{1}{c!}{\theta_1}^{i}{\theta_2}^{j} P_0
\end{aligned}$$ □

(8) For $c + 1 \leq i \leq c + m - 1$ and $j = 0$:

$$\begin{aligned}
P_{i,0} =& \frac{\lambda_1 P_{i-1,0} + c\mu P_{i+1,0} + c\mu P_{i,1}}{\lambda_1 + \lambda_2 + c\mu} \\
=& \frac{\frac{\lambda_1}{c^{i-1-c}c!}{\theta_1}^{i-1} + \frac{c\mu}{c^{i+1-c}c!}{\theta_1}^{i+1} + \frac{c\mu}{c^{i+1-c}c!}{\theta_1}^{i}\theta_2}{\lambda_1 + \lambda_2 + c\mu} P_0 \\
=& \frac{\frac{1}{c^{i-c}c!}{\theta_1}^{i}\left(\frac{c\lambda_1}{\theta_1} + \mu\theta_1 + \mu\theta_2\right)}{\lambda_1 + \lambda_2 + c\mu} P_0 \\
=& \frac{\frac{1}{c^{i-c}c!}{\theta_1}^{i}(c\mu + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + c\mu} P_0 \\
=& \frac{1}{c^{i-c}c!}{\theta_1}^{i} P_0
\end{aligned}$$ □

(9) For $c+1 \leq j \leq c+m-1$ and $i=0$:

$$\begin{aligned}
P_{0,j} =& \frac{\lambda_2 P_{0,j-1} + c\mu P_{1,j} + c\mu P_{0,j+1}}{\lambda_1+\lambda_2+c\mu} \\
=& \frac{\frac{\lambda_2}{c^{j-1-c}c!}{\theta_2}^{j-1} + \frac{c\mu}{c^{1+j-c}c!}\theta_1{\theta_2}^{j} + \frac{c\mu}{c^{j+1-c}c!}{\theta_2}^{j+1}}{\lambda_1+\lambda_2+c\mu} P_0 \\
=& \frac{\frac{1}{c^{j-c}c!}{\theta_2}^{j}\left(\frac{c\lambda_2}{\theta_2} + \mu\theta_1 + \mu\theta_2\right)}{\lambda_1+\lambda_2+c\mu} P_0 \\
=& \frac{\frac{1}{c^{j-c}c!}{\theta_2}^{j}(c\mu+\lambda_1+\lambda_2)}{\lambda_1+\lambda_2+c\mu} P_0 \\
=& \frac{1}{c^{j-c}c!}{\theta_2}^{j} P_0
\end{aligned}$$

□

(10) For $c+1 \leq i+j \leq c+m-1$ and $i,j \neq 0$:

$$\begin{aligned}
P_{i,j} =& \frac{\lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1} + c\mu P_{i+1,j} + c\mu P_{i,j+1}}{\lambda_1+\lambda_2+2c\mu} \\
=& \frac{\frac{\lambda_1}{c^{i-1+j-c}c!}{\theta_1}^{i-1}{\theta_2}^{j} + \frac{\lambda_2}{c^{i+j-1-c}c!}{\theta_1}^{i}{\theta_2}^{j-1} + \frac{c\mu}{c^{i+1+j-c}c!}{\theta_1}^{i+1}{\theta_2}^{j} + \frac{c\mu}{c^{i+j+1-c}c!}{\theta_1}^{i}{\theta_2}^{j+1}}{\lambda_1+\lambda_2+2c\mu} P_0 \\
=& \frac{\frac{1}{c^{i+j-c}c!}{\theta_1}^{i}{\theta_2}^{j}\left(\frac{c\lambda_1}{\theta_1} + \frac{c\lambda_2}{\theta_2} + \mu\theta_1 + \mu\theta_2\right)}{\lambda_1+\lambda_2+2c\mu} P_0 \\
=& \frac{\frac{1}{c^{i+j-c}c!}{\theta_1}^{i}{\theta_2}^{j}(c\mu+c\mu+\lambda_1+\lambda_2)}{\lambda_1+\lambda_2+2c\mu} P_0 \\
=& \frac{\frac{1}{c^{i+j-c}c!}{\theta_1}^{i}{\theta_2}^{j}(2c\mu+\lambda_1+\lambda_2)}{\lambda_1+\lambda_2+2c\mu} P_0 \\
=& \frac{1}{c^{i+j-c}c!}{\theta_1}^{i}{\theta_2}^{j} P_0
\end{aligned}$$

□

(11) For $i=0$ and $j=c+m$:

$$\begin{aligned}
P_{0,c+m} =& \frac{\lambda_2 P_{0,c+m-1}}{c\mu} \\
=& \frac{\lambda_2 \frac{1}{c^{c+m-1-c}c!}{\theta_2}^{c+m-1}}{c\mu} P_0 \\
=& \frac{\frac{1}{c^m c!}{\theta_2}^{c+m}\left(c\frac{\lambda_2}{\theta_2}\right)}{c\mu} P_0 \\
=& \frac{\frac{1}{c^m c!}{\theta_2}^{c+m} c\mu}{c\mu} P_0 \\
=& \frac{1}{c^m c!}{\theta_2}^{c+m} P_0
\end{aligned}$$

□

(12) For $i = c + m$ and $j = 0$:

$$\begin{aligned}
P_{c+m,0} =& \frac{\lambda_1 P_{c+m-1,0}}{c\mu} \\
=& \frac{\lambda_1 \frac{1}{c^{c+m-1-c}c!}{\theta_1}^{c+m-1}}{c\mu} P_0 \\
=& \frac{\frac{1}{c^m c!}{\theta_1}^{c+m}\left(c\frac{\lambda_1}{\theta_1}\right)}{c\mu} P_0 \\
=& \frac{\frac{1}{c^m c!}{\theta_1}^{c+m} c\mu}{c\mu} P_0 \\
=& \frac{1}{c^m c!}{\theta_1}^{c+m} P_0
\end{aligned}$$

□

(13) For $i + j = c + m$ and $i, j \neq 0$:

$$\begin{aligned}
P_{i,j} =& \frac{\lambda_1 P_{i-1,j} + \lambda_2 P_{i,j-1}}{2c\mu} \\
=& \frac{\lambda_1 \frac{1}{c^{i-1+j-c}c!}{\theta_1}^{i-1}{\theta_2}^{j} + \lambda_2 \frac{1}{c^{i+j-1-c}c!}{\theta_1}^{i}{\theta_2}^{j-1} +}{2c\mu} P_0 \\
=& \frac{\frac{1}{c^m c!}{\theta_1}^{i}{\theta_2}^{j}\left(\frac{c\lambda_1}{\theta_1} + \frac{c\lambda_2}{\theta_2}\right)}{2c\mu} P_0 \\
=& \frac{\frac{1}{c^m c!}{\theta_1}^{i}{\theta_2}^{j}(c\mu + c\mu)}{2c\mu} P_0 \\
=& \frac{\frac{1}{c^m c!}{\theta_1}^{i}{\theta_2}^{j}(2c\mu)}{2c\mu} P_0 \\
=& \frac{1}{c^m c!}{\theta_1}^{i}{\theta_2}^{j} P_0
\end{aligned}$$

□

Appendix B

Proof of Theorem 3.3.3

(1) For $n = 0$:

$$\begin{aligned} P_1 =& \frac{\lambda_1 + \lambda_2}{\mu} P_0 \\ =& \theta P_0 \end{aligned} \qquad \square$$

(2) For $1 \leq n \leq (k-1)$:

$$\begin{aligned} P_n =& \frac{(\lambda_1 + \lambda_2) P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1 + \lambda_2 + n\mu} \\ =& \frac{(\lambda_1 + \lambda_2) \frac{1}{(n-1)!} \theta^{n-1} P_0 + (n+1)\mu \frac{1}{(n+1)!} \theta^{n+1} P_0}{\lambda_1 + \lambda_2 + n\mu} \\ =& \frac{(\lambda_1 + \lambda_2) \frac{1}{(n-1)!} \left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n-1} P_0 + (n+1)\mu \frac{1}{(n+1)!} \left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n+1} P_0}{\lambda_1 + \lambda_2 + n\mu} \\ =& \frac{\frac{1}{n!} \left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n} P_0 (n\mu + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + n\mu} \\ =& \frac{1}{n!} \left(\frac{\lambda_1 + \lambda_2}{\mu}\right)^{n} P_0 \\ =& \frac{1}{n!} \theta^n P_0 \end{aligned} \qquad \square$$

(3) For $n = k$:

$$\begin{aligned}
P_n =& \frac{(\lambda_1+\lambda_2)P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1 + n\mu} \\
=& \frac{(\lambda_1+\lambda_2)\frac{1}{(n-1)!}\theta^{n-1}P_0 + (n+1)\mu\frac{1}{(n+1)!}{\theta_1}^{n+1-n}\theta^n P_0}{\lambda_1 + n\mu} \\
=& \frac{(\lambda_1+\lambda_2)\frac{1}{(n-1)!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n-1}P_0 + (n+1)\mu\frac{1}{(n+1)!}\left(\frac{\lambda_1}{\mu}\right)\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n}P_0}{\lambda_1 + n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n}P_0[n\mu+\lambda_1]}{\lambda_1 + n\mu} \\
=& \frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n}P_0 \\
=& \frac{1}{n!}\theta^n P_0
\end{aligned}$$

□

(4) For $(k+1) \leq n \leq c-1$:

$$\begin{aligned}
P_n =& \frac{\lambda_1 P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1 + n\mu} \\
=& \frac{\lambda_1\frac{1}{(n-1)!}{\theta_1}^{n-1-k}\theta^k P_0 + (n+1)\mu\frac{1}{(n+1)!}{\theta_1}^{n+1-k}\theta^k P_0}{\lambda_1 + n\mu} \\
=& \frac{\lambda_1\frac{1}{(n-1)!}\left(\frac{\lambda_1}{\mu}\right)^{n-1-k}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{k}P_0 + (n+1)\mu\frac{1}{(n+1)!}\left(\frac{\lambda_1}{\mu}\right)^{n+1-k}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{k}P_0}{\lambda_1 + n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1}{\mu}\right)^{n-k}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{k}P_0\left[\frac{n\lambda_1}{\frac{\lambda_1}{\mu}} + \frac{\lambda_1}{\mu}\mu\right]}{\lambda_1 + n\mu} \\
=& \frac{\frac{1}{n!}\left(\theta_1\right)^{n-k}\theta^k P_0\left[n\mu+\lambda_1\right]}{\lambda_1 + n\mu} \\
=& \frac{1}{n!}{\theta_1}^{n-k}\theta^k P_0
\end{aligned}$$

□

(5) For $n = c$:

$$\begin{aligned}
P_n =& \frac{\lambda_1 P_{n-1}}{n\mu} \\
=& \frac{\lambda_1 \frac{1}{(n-1)!} {\theta_1}^{n-k-1} \theta^k P_0}{n\mu} \\
=& \frac{1}{n\mu} \lambda_1 \frac{1}{(n-1)!} {\theta_1}^{n-k-1} \theta^k P_0 \\
=& \frac{\lambda_1}{\mu} \frac{1}{n(n-1)!} {\theta_1}^{n-k-1} \theta^k P_0 \\
=& \theta_1 \frac{1}{n!} {\theta_1}^{n-k-1} \theta^k P_0 \\
=& \frac{1}{n!} {\theta_1}^{n-k} \theta^k P_0 \qquad \square
\end{aligned} \tag{B.1}$$

Appendix C

Proof of Theorem 3.3.4

(1) For $n = 0$:

$$\begin{aligned} P_1 =& \frac{\lambda_1 + \lambda_2}{\mu} P_0 \\ =& \theta P_0 \end{aligned} \qquad \square$$

(2) For $1 \leq n \leq (a-1)$:

$$\begin{aligned} P_n =& \frac{(\lambda_1 + \lambda_2) P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1 + \lambda_2 + n\mu} \\ =& \frac{(\lambda_1 + \lambda_2) \frac{1}{(n-1)!} \theta^{n-1} P_0 + (n+1)\mu \frac{1}{(n+1)!} \theta^{n+1} P_0}{\lambda_1 + \lambda_2 + n\mu} \\ =& \frac{(\lambda_1 + \lambda_2) \frac{1}{(n-1)!} \left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n-1} P_0 + \mu \frac{1}{n!} \left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n+1} P_0}{\lambda_1 + \lambda_2 + n\mu} \\ =& \frac{\frac{1}{n!} \left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n} P_0 (n\mu + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + n\mu} \\ =& \frac{1}{n!} \theta^n P_0 \end{aligned} \qquad \square$$

(3) For $n = a$:

$$\begin{aligned}
P_n =& \frac{(\lambda_1+\lambda_2)P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1+\lambda_2{}' + n\mu} \\
=& \frac{(\lambda_1+\lambda_2)\frac{1}{(n-1)!}\theta^{n-1}P_0 + (n+1)\mu\frac{1}{(n+1)!}\theta^n(\theta_1+\theta_2')^{n+1-n}P_0}{\lambda_1+\lambda_2{}' + n\mu} \\
=& \frac{(\lambda_1+\lambda_2)\frac{1}{(n-1)!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n-1}P_0 + \mu\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n}\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)P_0}{\lambda_1+\lambda_2{}' + n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n}P_0\left(n\mu + \mu\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)\right)}{\lambda_1+\lambda_2{}' + n\mu} \\
=& \frac{\frac{1}{n!}\theta^n P_0(n\mu+\lambda_1+\lambda_2{}')}{\lambda_1+\lambda_2{}' + n\mu} \\
=& \frac{1}{n!}\theta^n P_0
\end{aligned}$$

□

(4) For $n = a + 1$:

$$\begin{aligned}
P_n =& \frac{(\lambda_1+\lambda_2{}')P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1+\lambda_2{}' + n\mu} \\
=& \frac{(\lambda_1+\lambda_2{}')\frac{1}{(n-1)!}\theta^{n-1}P_0 + (n+1)\mu\frac{1}{(n+1)!}\theta^a(\theta_1+\theta_2')^{n+1-a}P_0}{\lambda_1+\lambda_2{}' + n\mu} \\
=& \frac{(\lambda_1+\lambda_2{}')\frac{1}{(n-1)!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{a}P_0 + \mu\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{a}\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{n+1-a}P_0}{\lambda_1+\lambda_2{}' + n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{a}\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{n-a}P_0(n\mu+\lambda_1+\lambda_2{}')}{\lambda_1+\lambda_2{}' + n\mu} \\
=& \frac{1}{n!}\theta^a(\theta_1+\theta_2')^{n-a}P_0
\end{aligned}$$

□

(5) For $(a+1) < n \leq b$:

$$\begin{aligned}
P_n =& \frac{(\lambda_1 + \lambda_2{}')P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1 + \lambda_2{}' + n\mu} \\
=& \frac{(\lambda_1 + \lambda_2{}')\frac{1}{(n-1)!}\theta^a(\theta_1 + \theta_2')^{n+1-a}P_0 + (n+1)\mu\frac{1}{(n+1)!}\theta^a(\theta_1 + \theta_2')^{n+1-a}P_0}{\lambda_1 + \lambda_2{}' + n\mu} \\
=& \frac{(\lambda_1 + \lambda_2{}')\frac{1}{(n-1)!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^a\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{n-1-a}P_0 + \mu\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^a\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{n+1-a}P_0}{\lambda_1 + \lambda_2{}' + n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^n\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{n-a}P_0(n\mu + \lambda_1 + \lambda_2{}')}{\lambda_1 + \lambda_2{}' + n\mu} \\
=& \frac{1}{n!}\theta^a(\theta_1 + \theta_2')^{n-a}P_0
\end{aligned}$$

□

(6) For $n = b + 1$:

$$\begin{aligned}
P_n =& \frac{(\lambda_1 + \lambda_2')P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1 + \lambda_2 + n\mu} \\
=& \frac{\frac{(\lambda_1+\lambda_2')}{(n-1)!}\theta^a(\theta_1 + \theta_2')^{n-a-1}P_0 + \frac{(n+1)\mu}{(n+1)!}\theta^{n+1-b+a-1}(\theta_1 + \theta_2')^{b-a+1}P_0}{\lambda_1 + \lambda_2 + n\mu} \\
=& \frac{\frac{(\lambda_1+\lambda_2')}{(n-1)!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^a\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{n-a-1}P_0 + \frac{\mu}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{a+1}\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{n-a}P_0}{\lambda_1 + \lambda_2 + n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^a\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{n-a}P_0\left(n(\lambda_1 + \lambda_2')\left(\frac{\mu}{\lambda_1+\lambda_2'}\right) + \mu\left(\frac{\lambda_1+\lambda_2}{\mu}\right)\right)}{\lambda_1 + \lambda_2 + n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^a\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{n-a}P_0(n\mu + \lambda_1 + \lambda_2)}{\lambda_1 + \lambda_2 + n\mu} \\
=& \frac{1}{n!}\theta^a(\theta_1 + \theta_2')^{n-a}P_0
\end{aligned}$$

□

(7) For $(b+2) \leq n \leq (k-1)$:

$$\begin{aligned}
P_n =& \frac{(\lambda_1+\lambda_2)P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1+\lambda_2+n\mu} \\
=& \frac{\frac{(\lambda_1+\lambda_2)}{(n-1)!}\theta^{n-1-b+a-1}(\theta_1+\theta_2')^{b-a+1}P_0 + \frac{(n+1)\mu}{(n+1)!}\theta^{n+1-b+a-1}(\theta_1+\theta_2')^{b-a+1}P_0}{\lambda_1+\lambda_2+n\mu} \\
=& \frac{\frac{(\lambda_1+\lambda_2)}{(n-1)!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n-b+a-2}\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{b-a+1}P_0 + \frac{\mu}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n-b+a}\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{b-a+1}P_0}{\lambda_1+\lambda_2+n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n-b+a-1}\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{b-a+1}P_0(n\mu+\lambda_1+\lambda_2)}{\lambda_1+\lambda_2+n\mu} \\
=& \frac{1}{n!}\theta^{n-b+a-1}(\theta_1+\theta_2')^{b-a+1}P_0
\end{aligned}$$

□

(8) For $n = k$:

$$\begin{aligned}
P_n =& \frac{(\lambda_1+\lambda_2)P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1+n\mu} \\
=& \frac{\frac{(\lambda_1+\lambda_2)}{(n-1)!}\theta^{n-1-b+a-1}(\theta_1+\theta_2')^{b-a+1}P_0 + \frac{(n+1)\mu}{(n+1)!}\theta^{n-b+a-1}{\theta_1}^{n+1-n}(\theta_1+\theta_2')^{b-a+1}P_0}{\lambda_1+n\mu} \\
=& \frac{\frac{(\lambda_1+\lambda_2)}{(n-1)!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n-b+a-2}\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{b-a+1}P_0 + \frac{\mu}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n-b+a-1}\frac{\lambda_1}{\mu}\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{b-a+1}P_0}{\lambda_1+n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n-b+a-1}\left(\frac{\lambda_1+\lambda_2{}'}{\mu}\right)^{b-a+1}P_0\left(n\mu+\mu\left(\frac{\lambda_1}{\mu}\right)\right)}{\lambda_1+n\mu} \\
=& \frac{\frac{1}{n!}\theta^{n-b+a-1}(\theta_1+\theta_2')^{b-a+1}P_0(n\mu+\lambda_1)}{\lambda_1+n\mu} \\
=& \frac{1}{n!}\theta^{n-b+a-1}(\theta_1+\theta_2')^{b-a+1}P_0
\end{aligned}$$

□

(9) For $n = k + 1$:

$$\begin{aligned}
P_n =& \frac{\lambda_1 P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1 + n\mu} \\
=& \frac{\frac{\lambda_1}{(n-1)!}\theta^{n-1-b+a-1}(\theta_1+\theta_2')^{b-a+1}P_0 + \frac{(n+1)\mu}{(n+1)!}\theta^{k-b+a-1}{\theta_1}^{n+1-k}(\theta_1+\theta_2')^{b-a+1}P_0}{\lambda_1 + n\mu} \\
=& \frac{\frac{\lambda_1}{(n-1)!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{k-b+a-1}\left(\frac{\lambda_1+\lambda_2'}{\mu}\right)^{b-a+1}P_0}{\lambda_1 + n\mu} \\
&+ \frac{\frac{\mu}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{k-b+a-1}\left(\frac{\lambda_1}{\mu}\right)^{n+1-k}\left(\frac{\lambda_1+\lambda_2'}{\mu}\right)^{b-a+1}P_0}{\lambda_1 + n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{k-b+a-1}\left(\frac{\lambda_1}{\mu}\right)^{n-k}\left(\frac{\lambda_1+\lambda_2'}{\mu}\right)^{b-a+1}P_0\left(n\lambda_1\left(\frac{\mu}{\lambda_1}\right)+\mu\left(\frac{\lambda_1}{\mu}\right)\right)}{\lambda_1 + n\mu} \\
=& \frac{\frac{1}{n!}\theta^{k-b+a-1}{\theta_1}^{n-k}(\theta_1+\theta_2')^{b-a+1}P_0(n\mu+\lambda_1)}{\lambda_1 + n\mu} \\
=& \frac{1}{n!}\theta^{k-b+a-1}{\theta_1}^{n-k}(\theta_1+\theta_2')^{b-a+1}P_0
\end{aligned}$$

□

(10) For $(k+2) \leq n \leq (c-1)$:

$$\begin{aligned}
P_n =& \frac{\lambda_1 P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_1 + n\mu} \\
=& \frac{\frac{\lambda_1}{(n-1)!}\theta^{k-1-b+a}{\theta_1}^{n-1-k}(\theta_1+\theta_2')^{b-a+1}P_0 + \frac{(n+1)\mu}{(n+1)!}\theta^{k-b+a-1}{\theta_1}^{n+1-k}(\theta_1+\theta_2')^{b-a+1}P_0}{\lambda_1 + n\mu} \\
=& \frac{\frac{\lambda_1}{(n-1)!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{k-b+a-1}\left(\frac{\lambda_1}{\mu}\right)^{n-1-k}\left(\frac{\lambda_1+\lambda_2'}{\mu}\right)^{b-a+1}P_0}{\lambda_1 + n\mu} \\
&+ \frac{\frac{\mu}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{k-b+a-1}\left(\frac{\lambda_1}{\mu}\right)^{n+1-k}\left(\frac{\lambda_1+\lambda_2'}{\mu}\right)^{b-a+1}P_0}{\lambda_1 + n\mu} \\
=& \frac{\frac{1}{n!}\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{k-b+a-1}\left(\frac{\lambda_1}{\mu}\right)^{n-k}\left(\frac{\lambda_1+\lambda_2'}{\mu}\right)^{b-a+1}P_0\left(n\lambda_1\left(\frac{\mu}{\lambda_1}\right)+\mu\left(\frac{\lambda_1}{\mu}\right)\right)}{\lambda_1 + n\mu} \\
=& \frac{\frac{1}{n!}\theta^{k-b+a-1}{\theta_1}^{n-k}(\theta_1+\theta_2')^{b-a+1}P_0(n\mu+\lambda_1)}{\lambda_1 + n\mu} \\
=& \frac{1}{n!}\theta^{k-b+a-1}{\theta_1}^{n-k}(\theta_1+\theta_2')^{b-a+1}P_0
\end{aligned}$$

□

(11) For $n = c$:

$$\begin{aligned}
P_n =& \frac{\lambda_1 P_{n-1}}{n\mu} \\
=& \frac{\frac{\lambda_1}{(n-1)!}\theta^{k-1-b+a}{\theta_1}^{n-1-k}(\theta_1+\theta_2')^{b-a+1}P_0}{n\mu} \\
=& \frac{\lambda_1}{\mu}\frac{1}{n!}\theta^{k-b+a-1}{\theta_1}^{n-1-k}(\theta_1+\theta_2')^{b-a+1}P_0 \\
=& \theta_1\frac{1}{n!}\theta^{k-b+a-1}{\theta_1}^{n-k-1}(\theta_1+\theta_2')^{b-a+1}P_0 \\
=& \frac{1}{n!}\theta^{k-b+a-1}{\theta_1}^{n-k}(\theta_1+\theta_2')^{b-a+1}P_0 \qquad \square
\end{aligned} \tag{C.1}$$

Appendix D

Proof of Theorem 6.2.1

Due to the form of P_n various cases are considered. The required algebraic manipulation for each of the equations in 6.1 is given below:

(1) For $n = 0$:

$$\begin{aligned} P_0 =& \frac{\mu}{\lambda_a} P_1 \\ =& \frac{\mu}{\lambda_a} \left(\frac{1}{1!} {\theta_a}^1 P_0 \right) \\ =& \frac{\mu}{\lambda_a} \left(\frac{\lambda_a}{\mu} P_0 \right) \\ =& P_0 \end{aligned}$$ □

(2) For $1 \leq n \leq (k_1 - 1)$:

$$\begin{aligned} P_n =& \frac{\lambda_a P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_a + n\mu} \\ =& \frac{\lambda_a \frac{1}{(n-1)!} {\theta_a}^{n-1} P_0 + (n+1)\mu \frac{1}{(n+1)!} {\theta_a}^{n+1} P_0}{\lambda_a + n\mu} \\ =& \frac{\lambda_a \frac{1}{(n-1)!} \left(\frac{\lambda_a}{\mu} \right)^{n-1} P_0 + \mu \frac{1}{n!} \left(\frac{\lambda_a}{\mu} \right)^{n+1} P_0}{\lambda_a + n\mu} \\ =& \frac{\frac{1}{n!} \left(\frac{\lambda_a}{\mu} \right)^{n} P_0 (n\mu + \lambda_a)}{\lambda_a + n\mu} \\ =& \frac{1}{n!} {\theta_a}^n P_0 \end{aligned}$$ □

(3) For $n = k_1$:

$$\begin{aligned}
P_n =& \frac{\lambda_a P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_b + n\mu} \\
=& \frac{\lambda_a \frac{1}{(n-1)!} {\theta_a}^{n-1} P_0 + (n+1)\mu \frac{1}{(n+1)!} {\theta_a}^{k_1} {\theta_b}^{n+1-k_1} P_0}{\lambda_b + n\mu} \\
=& \frac{\frac{1}{n!} \left(\frac{\lambda_a}{\mu}\right)^n P_0 (n\mu + \lambda_b)}{\lambda_b + n\mu} \\
=& \frac{1}{n!} {\theta_a}^n P_0
\end{aligned}$$ □

(4) For $n = k_1 + 1$:

$$\begin{aligned}
P_n =& \frac{\lambda_b P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_b + n\mu} \\
=& \frac{\lambda_b \frac{1}{(n-1)!} {\theta_a}^{n-1} P_0 + (n+1)\mu \frac{1}{(n+1)!} {\theta_a}^{k_1} {\theta_b}^{n+1-k_1} P_0}{\lambda_b + n\mu} \\
=& \frac{\frac{1}{n!} {\theta_a}^{k_1} \theta_b P_0 \left(n\mu + \lambda_b\right)}{\lambda_b + n\mu} \\
=& \frac{1}{n!} {\theta_a}^{k_1} \theta_b P_0
\end{aligned}$$ □

(5) For $(k_1 + 1) \leq n \leq (k_2 - 1)$:

$$\begin{aligned}
P_n =& \frac{\lambda_b P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_b + n\mu} \\
=& \frac{\lambda_b \frac{1}{(n-1)!} {\theta_a}^{k_1} {\theta_b}^{n-1-k_1} P_0 + (n+1)\mu \frac{1}{(n+1)!} {\theta_a}^{k_1} {\theta_b}^{n+1-k_1} P_0}{\lambda_b + n\mu} \\
=& \frac{\frac{1}{n!} {\theta_a}^{k_1} {\theta_b}^{n-k_1} P_0 (n\mu + \lambda_b)}{\lambda_b + n\mu} \\
=& \frac{1}{n!} {\theta_a}^{k_1} {\theta_b}^{n-k_1} P_0
\end{aligned}$$ □

(6) For $n = k_2$:

$$\begin{aligned}
P_n =& \frac{\lambda_b P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_c + n\mu} \\
=& \frac{\lambda_b \frac{1}{(n-1)!} {\theta_a}^{k_1} {\theta_b}^{n-1-k_1} P_0 + (n+1)\mu \frac{1}{(n+1)!} {\theta_a}^{k_1} {\theta_b}^{k_2-k_1} {\theta_c}^{n+1-k_2} P_0}{\lambda_c + n\mu} \\
=& \frac{\frac{1}{n!} {\theta_a}^{k_1} {\theta_b}^{n-k_1} P_0 (n\mu + \lambda_c)}{\lambda_c + n\mu} \\
=& \frac{1}{n!} {\theta_a}^{k_1} {\theta_b}^{n-k_1} P_0
\end{aligned}$$ □

(7) For $n = k_2 + 1$:

$$\begin{aligned}
P_n =& \frac{\lambda_c P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_c + n\mu} \\
=& \frac{\lambda_c \frac{1}{(n-1)!} {\theta_a}^{k_1} {\theta_b}^{n-1-k_1} P_0 + (n+1)\mu \frac{1}{(n+1)!} {\theta_a}^{k_1} {\theta_b}^{k_2-k_1} {\theta_c}^{n+1-k_2} P_0}{\lambda_c + n\mu} \\
=& \frac{\frac{1}{n!} {\theta_a}^{k_1} {\theta_b}^{k_2-k_1} \theta_c P_0 (n\mu + \lambda_c)}{\lambda_c + n\mu} \\
=& \frac{1}{n!} {\theta_a}^{k_1} {\theta_b}^{k_2-k_1} \theta_c P_0
\end{aligned} \qquad \square$$

(8) For $(k_2 + 1) < n \leq (c - 1)$:

$$\begin{aligned}
P_n =& \frac{\lambda_c P_{n-1} + (n+1)\mu P_{n+1}}{\lambda_c + n\mu} \\
=& \frac{\lambda_c \frac{1}{(n-1)!} {\theta_a}^{k_1} {\theta_b}^{n-1-k_2} P_0 + (n+1)\mu \frac{1}{(n+1)!} {\theta_a}^{k_1} {\theta_b}^{k_2-k_1} {\theta_c}^{n+1-k_2} P_0}{\lambda_c + n\mu} \\
=& \frac{\frac{1}{n!} {\theta_a}^{k_1} {\theta_b}^{k_2-k_1} {\theta_c}^{n-k_2} P_0 (n\mu + \lambda_c)}{\lambda_c + n\mu} \\
=& \frac{1}{n!} {\theta_a}^{k_1} {\theta_b}^{k_2-k_1} {\theta_c}^{n-k_2} P_0
\end{aligned} \qquad \square$$

(9) For $n = c$:

$$\begin{aligned}
P_n =& \frac{\lambda_c P_{n-1}}{n\mu} \\
=& \frac{\frac{\lambda_c}{(n-1)!} {\theta_a}^{k_1} {\theta_b}^{k_2-k_1} {\theta_c}^{n-1-k_2} P_0}{n\mu} \\
=& \frac{1}{n!} {\theta_a}^{k_1} {\theta_b}^{k_2-k_1} {\theta_c}^{n-k_2} P_0
\end{aligned} \qquad \square \qquad \text{(D.1)}$$

Appendix E

Gaussian Elimination Algorithm

1. The reduction step:

$$\begin{aligned}
&\text{For } i = 1, 2, \ldots, (m+1) \times (n+1) - 1 : \\
&\quad a_{ji} = -\frac{a_{ji}}{a_{ii}} && \text{for all } j > i \\
&\quad a_{jk} = a_{jk} + a_{ji} a_{ik} && \text{for all } j, k > i
\end{aligned}$$

2. The back-substitution step:

$$\begin{aligned}
&x_t = 1 \\
&\text{For } i = t-1, t-2, \ldots, 1 \\
&\qquad x_i = -\frac{\left(\sum_{j=i+1}^{n} a_{ij} x_j \right)}{a_{ii}}
\end{aligned}$$

3. The normalisation step:

$$\begin{aligned}
&\text{norm} = \sum_{j=1}^{(m+1)\times(n+1)} x_j \\
&\text{For } i = 1, 2, \ldots, (m+1) \times (n+1) : \\
&\qquad \pi_i = \frac{x_i}{\text{norm}}
\end{aligned}$$

Bibliography

[1] I. Adan, G.-J. Houtum, and J. Wal. Upper and lower bounds for the waiting time in the symmetric shortest queue system. *Annals of Operations Research*, 48(2):197–217, April 1994.

[2] S. R. Agnihothri and P. F. Taylor. Staffing a centralized appointment scheduling department in Lourdes Hospital. *Interfaces*, 21(5):1–11, 1991.

[3] L. H. Aiken, S. P. Clarke, D. M. Sloane, J. Sochalski, and J. H. Silber. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of the American Medical Association*, 288(16):1987–93, 2012.

[4] G. Allon and A. Federgruen. Competition in service industries. *Operations Research*, 55(1):37–55, 2007.

[5] J. R. Artalejo and M. J. Lopez-Herrero. Analysis of the busy period for the M/M/c queue: An algorithmic approach. *Journal of applied probability*, 38:209–222, 2001.

[6] Audit Commission. The place of efficient and effective critical care services within the acute hospital. Technical report, 1999.

[7] Audit Commission. Bed Management. Technical report, 2003.

[8] R. J. Aumann. What is game theory trying to accomplish? In *Frontiers of Economics (K. J. Arrow and S. Honkapohja, eds.)*, pages 28–76. Basil Blackwell, Oxford, 1985.

[9] R. J. Aumann. Game theory. In Macmillan, editor, *The New Palgrave Volume 2 (J. Eatwell, M. Milgate, and P. Newman, Eds.)*, pages 460–482. London, 1987.

[10] N. T. J. Bailey. A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(2):185–199, 1952.

[11] N. T. J. Bailey. Queuing for medical care. *Applied Statistics*, 3(3):137–45, 1954.

[12] A. D. Banik and U. C. Gupta. Analyzing the finite buffer batch arrival queue under Markovian service process: GI(X)/MSP/1/N. *Top 15*, 1:146–160, 2007.

[13] R. Bekker and A. M. de Bruin. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, 178(1):45–65, June 2009.

[14] K. G. Binmore. Modeling rational players, Parts I and II. *Economics and Philosophy*, pages 179–214.

[15] E. L. Blair and C. E. Lawrence. A queueing network approach to health care planning with an application to burn care in New York State. *Socio-economic planning sciences*, 15(5):207–216, January 1981.

[16] M. A. Blegen, C. J. Goode, and L. Reed. Nurse staffing and patient outcomes. *Nursing Research*, 47(1):43–50, 1998.

[17] D. Braess. Uber ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung*, 12:258–268, 1968.

[18] D. Braess, A. Nagurney, and T. Wakolbinger. On a Paradox of Traffic Planning. *Transportation Science*, 39(4):446–450, November 2005.

[19] J. R. Broyles and J. K. Cochran. Estimating business loss to a hospital emergency department from patient reneging by queuing-based regression. In *Proceedings of the 2007 Industial Engineering Research Conference*, pages 613–618, 2007.

[20] G. P. Cachon and P. T. Harker. Competition and outsourcing with scale economies. *Management Science*, 48(10):1314–1333, 2002.

[21] G. P. Cachon and F. Zhang. Obtaining Fast Service in a Queueing System via Performance-Based Allocation of Demand. *Management Science*, 53(3):408–420, March 2007.

[22] W. Cahill and M. Render. Dynamic simulation modeling of ICU bed availability. *Simulation Conference Proceedings,*, pages 1573–1576, 1999.

[23] J. Caiado. Performance of combined double seasonal univariate time series models for forecasting water demand. *Journal of Hydrologic Engineering*, (March):215–222, 2009.

[24] N. Channouf, P. LEcuyer, A. Ingolfsson, and A. N. Avramidis. The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science*, 10(1):25–45, November 2006.

[25] M. L. Chaudhry, B. R. Madill, and G. Brière. Computational analysis of steady-state probabilities of M/G(a,b)/1 and related nonbulk queues. *Queueing Systems*, 2(2):93–114, 1987.

[26] T. J. Chaussalet, H. Xie, and P. Millard. A closed queueing network approach to the analysis of patient flow in health care systems. *Methods of Information in Medicine*, pages 492–497, 2006.

[27] S.-H. Cho, S. Ketefian, V. H. Barkauskas, and D. G. Smith. The effects of nurse staffing on adverse events, morbidity, mortality, and medical costs. *Nursing research*, 52(2):71–9, 2003.

[28] A. Chydzinski and R. Winiarczyk. On the blocking probability in batch Markovian arrival queues. *Microprocessors and Microsystems*, 32(1):45–52, 2008.

[29] J. Coast. Appropriateness versus efficiency: the economics of utilisation review. *Health policy (Amsterdam, Netherlands)*, 36(1):69–81, April 1996.

[30] T. J. Coats and S. Michalis. Mathematical modelling of patient flow through an accident and emergency department. *Emergency Medicine Journal*, 18(3):190–192, May 2001.

[31] J. K. Cochran and A. Bharti. Stochastic bed balancing of an obstetrics hospital. *Health Care Management Science*, 9(1):31–45, February 2006.

[32] J. K. Cochran and K. Roche. A queuing-based decision support methodology to estimate hospital inpatient bed demand. *Journal of the Operational Research Society*, 59(11):1471–1482, September 2008.

[33] M. Coombs and V. Lattimer. Safety, effectiveness and costs of different models of organising care for critically ill patients: literature review. *International journal of nursing studies*, 44(1):115–29, January 2007.

[34] J. K. Cooper and T. M. Corcoran. Estimating bed needs by means of queuing theory. *New England Journal of Medicine*, 291(8):404–405, 1974.

[35] A. X. Costa, S. A. Ridley, A. K. Shahani, P. R. Harper, V. De Senna, and M. S. Nielsen. Mathematical modelling and simulation for planning critical care capacity. *Anaesthesia*, 58(4):320–7, April 2003.

[36] A. A. Cournot. *Recherches sur les Principes Mathematiques de la Theorie des Richesses (English translation: Researches into the Mathematical Principles of the Theory of Wealth, New York: Macmillan, 1897.)*. Hachette, Paris, 1838.

[37] P. J. Courtois and J. Georges. On a Single-Server Finite Queuing Model with State-Dependent Arrival and Service Processes. *Operations Research*, 19(2):424–435, 1971.

[38] A. M. de Bruin, R. Bekker, L. Zanten, and G. M. Koole. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research*, 178(1):23–43, October 2009.

[39] A. M. de Bruin, G. M. Koole, and M. C. Visser. Bottleneck analysis of emergency cardiac in-patient flow in a university setting: an application of queueing theory. *Clinical and investigative medicine.*, 28(6):316–7, December 2005.

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

[41] S. Deo and I. Gurvich. Centralized vs. Decentralized Ambulance Diversion: A Network Perspective. *Management Science*, 57(7):1300–1319, May 2011.

[42] F. Downton. On limiting distributions arising in bulk service queues. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 18(2):265–274, 1956.

[43] P. Dubey. Inefficiency of Nash Equilibria. *Mathematics of Operations Research*, 11(1):1–8, 1986.

[44] G. J. Duke and J. V. Green. Night-shift discharge from intensive care unit increases the mortality-risk of ICU survivors. *Anaesthesia and intensive care*, 32(5):697–701, 2004.

[45] M. B. Dumas. Simulation modeling for hospital bed planning. *Simulation*, 43:69–78, 1984.

[46] D. L. Edbrooke, V. G. Stevens, C. L. Hibbert, A. J. Mann, and A. J. Wilson. A new method of accurately identifying costs of individual patients in intensive care: the initial results. *Intensive care medicine*, 23(6):645–50, June 1997.

[47] A. K. Erlang. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik, B*, 20:33, 1909.

[48] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektroteknikeren*, 10:189–197, 1917.

[49] Evolver an Official Website. http://www.palisade.com/evolver/. (Accessed on 01/02/2013).

[50] M. J. Faddy and S. I. McClean. Analysing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry*, 15(4):311–317, October 1999.

[51] Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt. Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science*, 54(2):324–338, February 2008.

[52] D. I. Fiems, G. Koole, and P. Nain. Waiting times of scheduled patients in the presence of emergency requests. *Technisch rapport. URL http://www.math.vu.nl/˜koole/ articles/ report05a/art.pdf, (Accessed on 18/12/2012)*, pages 1–19, 2007.

[53] S. Fomundam and J. W. Herrmann. A survey of queuing theory applications in healthcare. *Mechanical Engineering*, 2007.

[54] A. Fraser. *Computer Models in Genetics.* McGraw-Hill, New York, 1970.

[55] S. Gallivan and M. Utley. Devil's Advocacy and Patient Choice. In *29th Meeting of the EURO Working Group on Operational Research Applied to Health Services, Prague, Czech Republic*, pages 159–168, 2003.

[56] S. Gallivan and M. Utley. Modelling admissions booking of elective in-patients into a treatment centre. *IMA Journal of Management Mathematics*, 16(3):305–315, April 2005.

[57] S. Gallivan and M. Utley. A technical note concerning emergency bed demand. *Health care management science*, 14(3):250–2, September 2011.

[58] S. Gallivan, M. Utley, T. Treasure, and O. Valencia. Booked inpatient admissions and hospital capacity: mathematical modelling study. *British Medical Journal*, 324:280–282, 2002.

[59] D. P. Gaver. Markov chain analysis of a waiting-line process in continuous time. *The Annals of Mathematical Statistics*, 30(3):698–720, 1959.

[60] W. B. Gong, A. Yan, and C. G. Cassandras. The m / g / 1 queue with queue-length dependent arrival rate. 8(4):733–741, 1992.

[61] C. Goodeve. Operational research as a science. *Operations Research*, pages 166–181, 1953.

[62] F. Gorunescu, S. I. McClean, and P. H. Millard. Using a queueing model to help plan bed allocation in a department of geriatric medicine. *Health care management science*, 5(4):307–12, November 2002.

[63] W. K. Grassmann. Transient solutions in Markovian queueing systems. *Computers & Operations Research*, 4(1):47–53, 1977.

[64] L. Green. How many hospital beds? *Inquiry: a journal of medical care organization, provision and financing.*, 39(4):400–12, 2003.

[65] L. Green. Queueing analysis in healthcare. In *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 281–307. Springer, 2006.

[66] L. Green and P. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37(1):84–97, 1991.

[67] L. Green, P. Kolesar, and A. Svoronos. Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems. *Operations Research*, 39(3):502–511, May 1991.

[68] L. V. Green and P. J. Kolesar. The lagged PSA for estimating peak congestion in multiserver Markovian queues with periodic arrival rates. *Management Science*, 43(1):80–87, 1997.

[69] L. V. Green, P. J. Kolesar, and J. Soares. Improving the Sipp Approach for Staffing Service Systems That Have Cyclic Demands. *Operations Research*, 49(4):549–564, July 2001.

[70] L. V. Green and V. Nguyen. Strategies for cutting hospital beds: the impact on patient service. *Health services research*, 36(2):421–42, June 2001.

[71] L. V. Green, J. Soares, J. F. Giglio, and R. A. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*, 13(1):61–8, January 2006.

[72] J. D. Griffiths. Queueing at the Suez Canal. *Journal of the Operational Research Society*, 46:1299–1309, 1995.

[73] J. D. Griffiths and C. Cresswell. A mathematical model of a pelican crossing. *Journal of the Institute of Mathematics and Applications*, 18:381–394, 1976.

[74] J. D. Griffiths, V. Knight, and I. Komenda. Bed management in a Critical Care Unit. *IMA Journal of Management Mathematics*, 24(2):137–153, January 2013.

[75] J. D. Griffiths, N. Price-Lloyd, M. Smithies, and J. E. Williams. Modelling the requirement for supplementary nurses in an intensive care unit. *Journal of the Operational Research Society*, 56(2):126–133, November 2005.

[76] J.D. Griffiths, N. Price-Lloyd, M. Smithies, and J. Williams. A queueing model of activities in an intensive care unit. *IMA Journal of Management Mathematics*, 17(3):277, July 2006.

[77] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory (Wiley Series in Probability and Statistics)*. Wiley-Blackwell, 1998.

[78] R. Hagtvedt and M. Ferguson. Cooperative strategies to reduce ambulance diversion. In *Winter Simulation Conference*, pages 1861–1874, 2009.

[79] R. Hall, D. Belson, P. Murali, and M. Dessouky. Modeling patient flows through the healthcare system. In R.W. Hall, editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 1–44. Springer, 2006.

[80] P. R. Harper, J. Minty, B. Baffour, S. Sahu, and S. Sarran. MetSim: A Simulation Support Tool Using Meteorological Information to Improve the Planning and Management of Hospital Services. In *Proceedings of the 38th ORAHS Conference, Twente*, 2012.

[81] P. R. Harper, N. H. Powell, and J. E. Williams. Modelling the size and skill-mix of hospital nursing teams. *Journal of the Operational Research Society*, 61(5):768–779, May 2010.

[82] P. R. Harper and A. K. Shahani. Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research Society*, 53(1):11–18, January 2002.

[83] R. K. D. Haussmann. Waiting time as an index of quality of nursing care. *Health services research*, 5:92–105, 1970.

[84] A. G. Hawkes. Time-dependent solution of a priority queue with bulk arrival. *Operations Research*, 13(4):586–595, 1965.

[85] J. Healy, A. Thomas, J. Seargeant, and C. Victor. Coming Up for Care: Assessing the Post-Hospital Care Needs of Older People. *Policy Studies Institute, London*, 1999.

[86] J. C. Hershey, E. N. Weiss, and M. A. Cohen. A stochastic service network model with application to hospital facilities. *Operations Research*, 29(1):1–22, 1981.

[87] J. P. Holcomb and N. R. Sharpe. Forecasting police calls during peak times for the city of Cleveland. *Case Studies in Business, Industry, and Government Statistics*, 1(1):47–53, 2006.

[88] D. H. Howard. Why do transplant surgeons turn down organs? A model of the accept/reject decision. *Journal of health economics*, 21(6):957–69, November 2002.

[89] Aneurin Bevan Health Board Http://www.wales.nhs.uk/sitesplus/866/home. Aneurin Bevan Health Board - an Official NHS Wales website. (Accessed on 09/01/2013).

[90] A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li, and X. Wu. A Survey and Experimental Comparison of Service-Level-Approximation Methods for Nonstationary M(t)/M/s(t) Queueing Systems with Exhaustive Discipline. *INFORMS Journal on Computing*, 19(2):201–214, January 2007.

[91] N. Izady. *On Queues with Time-varying Demand*. PhD thesis, 2010.

[92] N. Izady and D. Worthington. Approximate analysis of non-stationary loss queues and networks of loss queues with general service time distributions. *European Journal of Operational Research*, 213(3):498–508, September 2011.

[93] N. Izady and D. Worthington. Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operational Research*, 219(3):531–540, June 2012.

[94] M. J. Jacob and T. P. Krishnamoorthy, A. Madhusoodanan. Transient solution of a finite-capacity M/Ga,B/1 queuing system. *Naval Research Logistics*, 35(3):437–441, 1988.

[95] I. W. Kabak. Blocking and delays in M(n)/M/C bulk queuing systems. *Operations Research*, 16(4):830–840, 1968.

[96] E. Kalai, M. I. Kamien, and M. Rubinovitch. Optimal Service Speeds in a Competitive Environment. *Management Science*, 38(8):1154–1163, August 1992.

[97] E. P. C. Kao and G. G. Tung. Bed allocation in a public health care delivery system. *Management Science*, 27(5), 1981.

[98] D. G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, 24(3):338–354, 1953.

[99] S. C. Kim and I. Horowitz. Scheduling hospital services : the efficacy of elective-surgery quotas. 30:335–346, 2002.

[100] S. C. Kim, I. Horowitz, and K. K. Young. Analysis of capacity management of the intensive care unit in a hospital. *European Journal of*, 115:36–46, 1999.

[101] S. C. Kim, I. Horowitz, K. K. Young, and T. A. Buckley. Flexible bed allocation and performance in the intensive care unit. *Journal of Operations Management*, 18(4):427–443, June 2000.

[102] V. A. Knight. A game theoretical approach to the Emergency Medical Vehicle - Emergency Department interface. *In Preparation*, 2013.

[103] V. A. Knight and P. Harper. Selfish Routing in Public Services. *European Journal of Operational Research*, (In Press), 2013.

[104] N. Koizumi, E. Kuno, and T. E. Smith. Modeling patient flows using a queuing network with blocking. *Health care management science*, 8(1):49–60, February 2005.

[105] P. J. Kolesar and K. L. Rider. A queuing-linear programming approach to scheduling police patrol cars. *Operations Research*, 23(6):1045–1062, 1975.

[106] A. Kolker. Process Modeling of ICU Patient Flow: Effect of Daily Load Leveling of Elective Surgeries on ICU Diversion. *Journal of Medical Systems*, 33(1):27–40, May 2008.

[107] I Komenda, J. D. Griffiths, and V. A. Knight. A model of CCU activities through queueing theory. In *Proceedings of ORAHS Conference*, 2012.

[108] E. Koutsoupias and C. Papadimitriou. Worst-Case Equilibria. In *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science*, pages 404–413, 1999.

[109] C. Kovner, C. Jones, C. Zhan, P. J. Gergen, and J. Basu. Nurse staffing and postsurgical adverse events: an analysis of administrative data from a sample of US hospitals, 1990-1996. *Health Services Research*, 37(3):611–629, 2002.

[110] D. M. Kreps. *Game theory and economic modelling*. Clarendon Press, Oxford, 1990.

[111] D. Levhari and I. Luski. Duopoly pricing and waiting lines. *European Economic Review*, 11:17–35, 1978.

[112] S. C. Lim, V. Doshi, B. Castasus, J. K. H. Lim, and K. Mamun. Factors causing delay in discharge of elderly patients in an acute care hospital. *Annals of the Academy of Medicine, Singapore*, 35(1):27–32, January 2006.

[113] N. Litvak, M. Vanrijsbergen, R. Boucherie, and M. Vanhoudenhoven. Managing the overflow of intensive care patients. *European Journal of Operational Research*, 185(3):998–1010, March 2008.

[114] E. M. Lopezsoriano, M. E. Matthews, and J. P. Norback. Improving the flow of customers in a hospital cafeteria. *Journal of the American Dietetic Association*, 79(6):683–688, 1981.

[115] R. D. Luce and H. Raiffa. *Games and decisions*. John Wiley and Sons, New York, 1957.

[116] J. M. C. Maessen, C. H. C. Dejong, A. G. H. Kessels, and M. F. von Meyenfeldt. Length of stay: an inappropriate readout of the success of enhanced recovery programs. *World journal of surgery*, 32(6):971–5, June 2008.

[117] F. Mallor and C. Azcárate. Combining optimization with simulation to obtain credible models for intensive care units. *Annals of Operations Research*, (Law 2006), December 2011.

[118] A. H. Marshall and S. I. McClean. Conditional phase-type distributions for modelling patient length of stay in hospital. *International Transactions in Operational Research*, 10(6):565–576, 2003.

[119] W. A. Massey and W. Whitt. A stochastic model to capture space and time dynamics in wireless communication systems. *Probability in the Engineering and Informational Sciences*, 8:541–569, 1994.

[120] W. A. Massey and W. Whitt. Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems*, 25:157–172, 1997.

[121] J. O. McClain. Bed planning using queuing theory models of hospital occupancy: a sensitivity analysis. *Inquiry*, 13(2):167–176, 1976.

[122] M. L. McManus, M. C. Long, A. Cooper, and E. Litvak. Queuing theory accurately models the need for critical care resources. *Anesthesiology*, 100(5):1271–6, May 2004.

[123] D. G. McQuarrie. Hospital utilization levels. The application of queuing theory to a controversial medical economic problem. *Minnesota Medicine*, 66(11):679–686, 1983.

[124] Merlin. Cardiff University's High Performance Computer. *http://www.cardiff.ac.uk/ arcca/ services/equipment/Software/merlin_software.html*, (Accessed on 29/03/2013).

[125] R. A. Milliken, L. Rosenberg, and G. M. Milliken. A queuing theory model for the prediction of delivery room utilization. *American Journal of Obstetrics and Gynecology*, 114(5):691–699, 1972.

[126] I. Mitchell and M. Grounds. Intensive care in the ailing UK health care system. *Lancet*, pages 1970–1970, 1995.

[127] J. F. Nash. Equilibrium Points in N-Person Games. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 48–49, 1950.

[128] J. Needleman and P. Buerhaus. Nurse-staffing levels and the quality of care in hospitals. *The New England journal of medicine*, 346(22):1715–22, May 2002.

[129] NHS Trust and PCT Combined Reference Cost Schedules. DOH. *http://www.dh.gov.uk/en/ Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_062884*, (Accessed on: 27/02/2012), 2006.

[130] NHS Trust and PCT Combined Reference Cost Schedules. DOH. *http://www.dh.gov.uk/en/ Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_082571*, (Accessed on: 27/02/2012), 2007.

[131] R. A. Nosek and J. P. Wilson. Queuing theory and customer satisfaction: a Review of terminology, trends, and applications to pharmacy practice. *Hospital Pharmacy*, 36(3):275–279, 2001.

[132] Office for National Statistics. Population Ageing in the United Kingdom, its Constituent Countries and the European Union. (March):1–12, 2012.

[133] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. Cambridge: MIT Press, 1994.

[134] R. Pate. What is Payment by Results? Technical Report May, 2009.

[135] F. Pollaczek. Uber das Warteproblem. *Mathematische Zeitschrift*, 38(1):492–537, 1934.

[136] J. Preater. Queues in health. *Health care management science*, 5(4):283, November 2002.

[137] A. E. Raftery. Bayesian model selection in social research. *Sociological methodology*, 25:111–163, 1995.

[138] J. C. Ridge, S. K. Jones, M. S. Nielsen, and A. K. Shahani. Capacity planning for intensive care units. *European journal of operational research*, 105(2):346–355, March 1998.

[139] K. T. Roche, J. K. Cochran, and I. A. Fulton. Improving patient safety by maximizing fast-track benefits in the emergency department a queuing network approach. In *Proceedings of the 2007 Industrial Engineering Research Conference*, pages 619–624, 2007.

[140] M. Rosen and Joint Working Party On Graduated Patient Care Royal College Of Anaesthetists. *Report of the Joint Working Party On Graduated Patient Care*. RCA & RCSE, 1996.

[141] C. J. Rosenquist. Queueing analysis: a useful planning and management technique for radiology. *Journal of Medical Systems*, 11(6):413–419, 1987.

[142] A. E. Roth. The evolution of the labor market for medical interns and residents: a case study in game theory. *The Journal of Political Economy*, 92(6):991–1016, 1984.

[143] M. B. Rothberg, I. Abraham, P. K. Lindenauer, and D. N. Rose. Improving nurse-to-patient staffing ratios as a cost-effective safety intervention. *Medical care*, 43(8):785–91, August 2005.

[144] T. Roughgarden. *Selfish Routing and the Price of Anarchy: Books*. MIT Press, 2005.

[145] T. Sasichay-Akkadechanunt, C. C. Scalzi, and A. F. Jawad. The relationship between nurse staffing and patient outcomes. *The Journal of nursing administration*, 33(9):478–85, September 2003.

[146] T. C. Schelling. *The strategy of conflict*. Harvard University Press, Cambridge, Mass., 1960.

[147] C. V. Seshaiah and H. B. Thiagaraj. A Queueing Network Congestion Model in Hospitals. *European journal of scientific research*, 63(3):419–427, 2011.

[148] A. K. Shahani, S. A. Ridley, and M. S. Nielsen. Modelling patient flows as an aid to decision making for critical care capacities and organisation. *Anaesthesia*, 63(10):1074–80, October 2008.

[149] A. Shmueli, C. L. Sprung, and E. H. Kaplan. Optimizing admissions to an intensive care unit. *Health care management science*, 6(3):131–6, August 2003.

[150] K. Siddhartan, W. J. Jones, and J. A. Johnson. A priority queuing model to reduce waiting times in emergency care. *International Journal of Health Care Quality Assurance*, 9(5):10–16, 1996.

[151] G. B. Smith, B. L. Taylor, P. J. McQuillan, and E. Nials. Rationing intensive care. Intensive care provision varies widely in Britain. *British Medical Journal*, 310(5):1412–1413, 1995.

[152] W. J. Stewart. *Probability, Markov Chains, Queues, and Simulation*. Priceton University Press, 1st edition, 2009.

[153] A. L. Stolyar. Optimal routing in output-queued flexible server systems. *Probability in the Engineering and Informational Sciences*, 19:141–189, 2005.

[154] M. Sznajder, G. Leleu, G. Buonamico, B. Auvert, P. Aegerter, Y. Merlière, M. Dutheil, B. Guidet, and J. R. Le Gall. Estimation of direct cost and resource allocation in intensive care: correlation with Omega system. *Intensive care medicine*, 24(6):582–9, June 1998.

[155] T. H. Taylor, A. M. C. Jennings, D. A. Nightingale, B. Barber, D. Leivers, M. Styles, and J. Magner. A study of anaesthetic emergency work. Paper 1: The method of study and introduction of queuing theory. *British Journal of Anaesthesia*, 41(1):70–75, 1969.

[156] T. Tezcan. Optimal Control of Distributed Parallel Server Systems Under the Halfin and Whitt Regime. *Mathematics of Operations Research*, 33(1):51–90, February 2008.

[157] S. Tuft and S. Gallivan. Computer modelling of a cataract waiting list. *The British journal of ophthalmology*, 85(5):582–5, May 2001.

[158] M. Utley and D. Worthington. *Handbook of Healthcare System Scheduling, Chapter 2: Capacity Planning*. 2011.

[159] A. Van Ackere. Conflicting Interests in the Timing of Jobs. *Management Science*, 36(8):970–984, 1990.

[160] G. Vassilacopoulos. Allocating doctors to shifts in an accident and emergency department. *Journal of the Operational Research Society*, 36(6):517–523, 1985.

[161] C. R. Victor, J. Healy, A. Thomas, and J. Seargeant. Older patients and delayed discharge from hospital. *Health & social care in the community*, 8(6):443–452, November 2000.

[162] C. R. Victor, B. Nazareth, and M. Hudson. The inappropriate use of acute beds in an inner London DHA. *Health Trends*, 25(3):94–97, 1993.

[163] J. Vile. *Time-Dependent Stochastic Modelling For Predicting Demand And Scheduling Of Emergency Medical Services*. PhD thesis, Cardiff University, 2012.

[164] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[165] A. D. Wall and D. J. Worthington. Time-dependent analysis of virtual waiting time behaviour in discrete time queues. *European Journal of Operational Research*, 178(2):482–499, April 2007.

[166] W. Whitt. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics (NRL)*, 54(5):476–484, 2007.

[167] T. Williams and G. Leslie. Delayed discharges from an adult intensive care unit. *Australian health review : a publication of the Australian Hospital Association*, 28(1):87–96, January 2004.

[168] W. L. Winston. *Operations Research: Applications and Algorithms*. Brooks/Cole, 1998.

[169] D. Worthington. Queueing models for hospital waiting lists. *Journal of the Operational Research Society*, 38(5):413–422, 1987.

[170] D. Worthington. Hospital waiting list management models. *Journal of the Operational Research Society*, 42(10):833–843, 1991.

[171] J. P. Young. Estimating bed requirements, in A Queuing theory approach to the control of hospital inpatient census. Technical report, John Hopkins University, Baltimore, 1962.

[172] J. P. Young. The basic models, in A Queuing theory approach to the control of hospital inpatient census. Technical report, John Hopkins University, Baltimore, 1962.

[173] J. P. Young. Stabilisation of inpatient bed occupancy through control of admissions. *Hospitals*, 39(19):41–48, 1965.